2015

Safe City Video Analytics Technologies

Homeland Security Research Corp.

Safe City Video Analytics Technologies

August 2015

Homeland Security Research Corp. (HSRC) *is an international market and technology research firm specializing in the Homeland Security (HLS) & Public Safety (PS) Industry. HSRC provides premium market reports on present and emerging technologies and industry expertise, enabling global clients to gain time-critical insight into business opportunities. HSRC's clients include U.S. Congress, DHS, U.S. Army, U.S. Navy, NATO, DOD, DOT, GAO, and EU, among others; as well as HLS & PS government agencies in Japan, Korea, Taiwan, Israel, Canada, UK, Germany, Australia, Sweden, Finland, Singapore. With over 750 private sector clients (72% repeat customers), including major defense and security contractors, and Fortune 500 companies. HSRC earned the reputation as the industry's Gold Standard for HLS & PS market reports.*

Washington D.C. 20004, 601 Pennsylvania Ave., NW Suite 900,
Tel: 202-455-0966, info@hsrc.biz, www.homelandsecurityresearch.com

Table of Contents

List of Tables

List of Figures

1 Safe City Video Analytics Technologies

1.1 Introduction

- Video analytics is defined as software (or firmware) that imitates the analysis that an operator would do looking at the video images from surveillance video cameras. The analytics software processes video stream images to automatically detect items (e.g. people) in security related events and commercial purposes. Once detected, the items can be identified, monitored and located. Their actions and interactions are analyzed and classified in order to interpret the activity of a scene and bring it to the attention of the operator.
- Video analytics and video content analysis software applications are used in two modes: real and non-real time. In real time, video analytics software detects situations in the video stream that represent a security threat and trigger an alarm. In non-real time, they make it possible to find video images on an incident under surveillance.
- Video analytic applications are divided into image analysis and application-specific parts. The interface between these two parts produces an abstraction that describes the scene based on the objects present. The application specific part performs a comparison of the scene descriptions and of the scene rules (such as virtual lines that are prohibited to cross, or polygons that define a protected area). Other rules may represent intra-object behavior such as objects following other objects (to form a tailgating detection). Such rules can also be used to describe prohibited object motion which may be used to establish a speed limit.
- Safe City Video content analysis (VCA) is the capability of automatically analyzing video to detect and determine temporal events not based on a single image. As such, it can be seen as the automated equivalent of human image interpretation. This technical capability is used in a wide range of domains including public safety, national security, entertainment, health-care, retail, transport, residential security, safety and general security.
- Safe city Intelligent Video Surveillance systems typically use a large number of video cameras, transmitting the video signals to a central command and control center, where a multiplex matrix is used to display certain images to security personnel.
- Event detection and recognition requires the perceptual capabilities of human operators to detect and identify items moving within the field-of-view of the video cameras and to understand their actions.

- No matter how vigilant the operators are, manual monitoring inevitably suffers from data overload as a result of periods of operator inattention due to fatigue, distractions and interruptions.
- In practice, it is inevitable that a significant proportion of the video channels are not regularly monitored and potentially important events are overlooked. Furthermore, fatigue increases radically as the number of video cameras in the system is increased. Automating all or part of this process would provide significant benefits ranging from the capability to alert an operator to potential events of interest, through to a fully automatic detection and analysis system. However, the reliability of automated detection systems is a very important issue since frequent false alarms induce doubt in the operators who quickly learn to ignore the system.

Table 1 - Safe City Video Analytics, Functions and Description

Function	Description
Dynamic masking	Blocking a part of the video signal based on the signal itself (e.g., privacy concerns).
Egomotion estimation	Egomotion estimation is used to determine the location of a camera by analyzing its output signal.
Motion detection	Motion detection is used to determine the presence of relevant motion in the observed scene.
Object detection	Object detection is used to determine the presence of a type of object or entity, such as a person or car. Other examples include fire and smoke detection.
Recognition	Face recognition and Automatic Number Plate Recognition are used to recognize, and therefore possibly identify persons or cars.
Style detection	Style detection is used in settings where the video signal has been produced. Style detection detects the style of the production process
Tamper detection	Tamper detection is used to determine whether the camera or its signal is tampered with.
Video tracking	Video tracking is used to determine the location of persons or objects in the video signal, possibly with regard to an external reference grid.

1.1.1 Video Analytics System Architecture

- Safe City video analytics architecture consists of elements and interfaces. Each element provides a functionality corresponding to a semantically unique entity of the complete video analytics solution. Interfaces are unidirectional and define an information entity with a unique content.
- Only the interfaces are subject to this specification. Central to this architecture is the ability to distribute any elements or sets of adjacent elements to any device in the network.

1.1.2 Intelligent Video Surveillance: Cloud Platforms

- Cloud platforms are becoming very popular for processing and storage for large scale safe cities video surveillance networks. In this context, privacy challenges arise which are associated with retention and dissemination of personal information. Then the key question is whether the current cloud platforms are indeed providing enough security standards and safeguards to ensure the protection of sensitive information. For cloud computing applications, destruction of data is not simple. After personal data is no longer useful, they are usually discarded. But to completely erase the collected and stored data is not an easy process since the information controller has to completely trust their provider to erase/remove the information when requested. Thus the security/privacy guidelines with respect to personal data are to establish mechanisms to autonomously (or after request) destruct personal information while at the same time, these mechanisms provide the capability to eliminate or minimize the risk of unauthorized data access. The requirement levels and features include (but are not limited to):
  - Data retention to determine where end-users have stored their personal information in the cloud hosting proxy
  - Data privacy regulations and legislations to determine a consistent data destruction policy which will have to be followed for all stored personal records
  - Data destruction for permanently deleting end-users' sensitive personal information.
- In fact, a number of IT processes can be applied involving writing over data files with other type of files containing junk information or involving re-alignment of the magnetic fields of storage devices. Another important issue is large databases which are directly related to the storage components.
- Databases need to be maintained in private premises and only be accessed via private networks called Private Cloud Platforms/Infrastructures. Furthermore, they need to be secured and auditable and thus there is a need to create reliable security, trust and advanced authentication / encryption mechanisms. Moreover, databases need to have dynamic data encryption levels with associated rules and mechanisms for dynamic setup of large, secure databases. Finally, all important private information must be segregated and reported properly in a highly sensitive fashion. This can be implemented with appropriate personal data management policies and standardized formats to respect and protect their sensitive content.
- In conclusion, for safe city situations, large scale scalable video surveillance networks comprised of hundreds or thousands of sensor

nodes using advanced machine intelligence techniques for better decision support for the end-user in–the loop operators. The involved data volumes require different levels of processing and different artificial intelligence techniques for reliable and autonomous pattern matching and crowed behavior analysis and prediction. The application fields are many and include a multitude of security - safety scenarios and test cases. In addition, due to the enormous computing and storage requirements, the cloud computational platforms are used in conjunction with the Internet of Things (IOT) emergent applications where real world devices are connected to the internet infrastructure. These platforms create by default new privacy and security challenges concerning sensitive data namely residence or various geographical and positional / geo-location recordings (such as GPS localization measurements and individual day to day behavior analysis and studies). The main considerations that need to be addressed include:

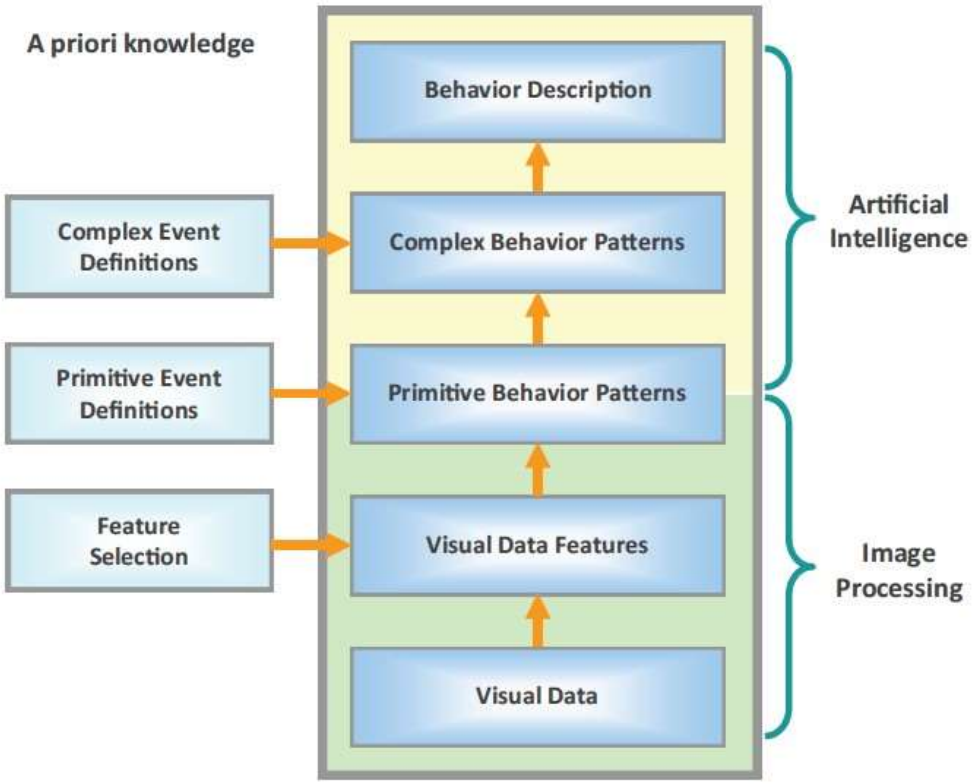
- Physical system protection of sensitive data
- Secure physical location of servers that store sensitive databases with high security and authorized access levels
- Dynamic set ups of security levels implemented where database access storage or transmission is required in conjunction
- Development of robust and immune to attacks encryption tools and algorithms for highly secure storage information and trusted data transmission.

1.2 Safe City Video Analytics Based Suspect Behavioral Analysis

- Terminology disambiguation: Behavior analysis is a term widely used by psychologists when referring to the foundations and principles of behaviorism. In this document, the term behavior analysis will be referred to as the process of tracking, modeling and analyzing behavior for surveillance purposes.
- Visual surveillance in dynamic scenes using multiple cameras attempts to detect, recognize and track certain objects from image sequences, and more importantly to understand and describe object behaviors. This type of surveillance technology, especially for humans, is currently one of the most active research topics in computer vision and artificial intelligence. It has a wide spectrum of public safety and security applications including access control, crowd flux statistics and congestion analysis as well as human behavior detection and analysis.
- Behavioral analysis typically requires video as input. This input can be obtained by a variety of means. Current off-the-shelf video surveillance

systems process video sequences and perform well in low level image processing functions such as object detection, recognition and tracking as well as object classification. Until recently, reliable object classification and understanding of human activities has been carried out manually by human operators. However, passive video surveillance is proving ineffective as the number of cameras exceed the capability of human operators to monitor them. The desired automation of the process is an open research question that involves the combination of image processing functions with artificial intelligence techniques. The goal of visual surveillance is not only to put cameras in the place of human eyes, but also to accomplish the behavior analysis task as automatically as possible.

Figure 1 - Architecture of a Safe City Video Understanding System



- In order to replace human operators and achieve automated behavior analysis, a more complex scene analysis is required. Behavior analysis involves detection and monitoring movement of identified objects (usually humans) by recognizing and classifying patterns of activities. The classification of human activities is not a trivial task due to the randomness and complex nature of human movement. The idea is to partition the observed human movements into some discrete states and

then classify them appropriately. Partitioning of the observed movements is very application-specific. Even with the use of context specific rules, it has been proven very hard to predict what constitutes suspicious or endangering behavior. Typical techniques adopted in behavioral analysis include: trajectory estimation, individual posture recognition, facial recognition, skeleton fitting and interaction modeling. The fusion of the various modalities (face, posture, trajectory etc.) is referred to as multimodal behavior analysis and allows for deeper modeling of human behavior.

- In the process of fusing various modalities, each media stream is associated with a different assurance level for a given surveillance task. For example, the system designer may have a higher poise in the video stream balanced to the audio stream for detecting human running events. The assurance level of streams is usually pre-calculated based on their earlier accuracy. This traditional approach is difficult especially when we insert a new stream in the system with no knowledge of its prior history.
- Typical limitations in behavior analysis are the varying camera angles and the process of camera handoff, the variation in environment lighting, the suspects' use of disguises and the alarmingly increasing volume of data collected.
- Behavior analysis is only involved in sophisticated versions of visual analysis. Stakeholders involved are typically security and defense agencies as well as other government authorities that wish to protect their citizens from terrorism and crime.

1.3 Video Surveillance as a Service (VSaaS)

- The recently emergent Video Surveillance as a Service (VSaaS) technological trend is enjoying some substantial and dynamic popularity and it is believed that it will continue to enjoy serious growth rates in the global consumer market especially due to the relevant cloud - application service providers.
- By bringing Software as a Service (SaaS) into physical security, hosted & managed video has the potential to make video surveillance cheaper and easier to deploy and use.
- More specifically, in the short term we expect managed & hosted video to:
  - Re-shape the home and small business market segments for video surveillance
  - Force traditional video surveillance providers to enhance their product offerings

- This is a fast changing market segment with new entrants joining the market every month and dozens of startups around the world.
- It is surprising to see the level of diversity among providers, with companies not only targeting different markets but fundamentally different architectures and widely varying pricing.
- The following differentiators should be considered:
  - Camera Support
  - On-Site Setup Complexity
  - Channel Partner / Strategy
  - Video Management Sophistication
  - Market Segment Targeted
  - Local Storage Support
  - Hosting Scalability
  - Pricing

1.3.1 Video Surveillance as Service Solutions: 26 Vendors

The following vendors provide diverse Video Surveillance as service solutions:

1. Archerfish
2. ByRemote
3. Alarm.com
4. Axis AVHS
5. Brivo
6. CameraManager
7. Connexed
8. Dropcam
9. DvTel
10. Envysion
11. EMC VSaaS Service
12. iControl
13. ipConfigure
14. Lorex (Tested)

15. Napco iSee Video
16. Navco Videometrixs
17. NeoVSP
18. OzVision
19. Rogo (Tested)
20. Sensr.net
21. Secure-i (Tested)
22. Starvedia (Tested)
23. Viaas (Tested)
24. VideoCells
25. Vue
26. Xanboo

1.4 Real Time Automatic Alerts Software

- Most item alerts are defined by the Intelligent Video Surveillance Safe City user. They may be generic alerts, for example detecting an abandoned item or an item in the scene moving over a set speed limit. To trigger these alarms, only the properties of the item movements are analyzed by the system. More obvious alerts may be issued after the items or their movement have been classified (e.g., discrimination between the passage of a human or animal in an outside area). Behavior-related alerts based on conformity or non-conformity with a behavior model entered in the system (e.g., an individual trying to open more than one car in a parking lot), constitute pre-defined alerts.
- Certain real-time alerts are automatically identified. Over time, the system learns a model of activity and ends up detecting non-standard activities. For example, analytics software may learn that vehicles drive on the street and pedestrians walk on the sidewalk. The opposite may trigger an alarm.
- Video search for surveillance: Analytics processing makes it possible to index video content based on signatures, for example the shape of items, their size, appearance, trajectory, type, as well as their model of activity. Stored as metadata, this data makes it possible to conduct spatio-temporal searches.
- The software developed for Intelligent Video Surveillance systems have different levels of analysis. Hierarchically, they are executed at the pixel and item level to achieve the behavior scale. They are grouped according to the following tasks:

- Sorting of moving items
- Monitoring of items
- Classification and ID of items
- Classification of activities and behaviors
- Detection of changes

- In video surveillance, detecting changes in video images is the basis for all intelligent analysis. It may detect an activity in a scene under surveillance, in particular the movement of items. It may also reveal the appearance or disappearance of an item (e.g., abandoned item). It is also used to automatically report accidental or intentional alterations in a camera: obstructions (e.g., dust, moisture, paint, and stickers), reorientation, and blurriness.
- Several software for detecting movements used in video analytics are based on detecting changes. However, detecting changes in video images does not clearly target the movement of items but may highlight an image modulation. In order to segment moving items, we must be able to discriminate between fluctuations in pixel value corresponding to consistent movements and fluctuations caused by background changes.

1.5 Image Segmentation Software

- Image segmentation is a key problem in video analytics because complex backgrounds may at certain times present many sudden variations such as change in illumination (shadows, movement of light source, clouds, light reflections, glare caused by sources of light) and other non-relevant movements. That is why many analytical methods work well indoors and in scenes with little movement. Software that is robust enough to be applied in uncontrolled settings is much rarer.
- Several movement segmentation softwares are used.
    - Subtraction of the background: In its most primitive form, detecting changes comes down to finding pixel by pixel, differences in color or texture between the images of a sequence. A first category of software consists in comparing each frame of a sequence to a reference image called the background which represents the undisturbed scene. The areas of change are formed of pixels with a difference in intensity that is above a threshold. Pixel-by-pixel subtraction between two images is very sensitive to the slightest background change, for example changes in illumination and movements inherent to a scene (e.g., shadow of a tree blowing in the wind). In order to offset this problem, certain software continually adapts the background model to intrinsic changes in the

background. The difference with the background is a technique that is particularly suited to indoor backgrounds where illumination conditions are controlled and where there is little activity (e.g., monitoring a hallway).

- Time-based difference: A second class of methods for detecting change is based on a time difference between a few consecutive video frames. These video frames adapt to variations in the time of the background. On the other hand, they tend to oversee certain variations related to the movement of items in the scene, especially if they move slowly. They often produce holes in the items detected. This software therefore requires smoothing treatment with morphological operators and filtering of holes and shapes that are too small. In order to retain only significant movements and eliminate occasional movements, certain softwares draw up a map of the regions with a high level of activity based on a movement pattern.
- Optical Flow: Methods that analyze optical flow help to detect consistent directions of pixel change associated with the movement of items in the scene. However, they require complex calculations that are difficult to do in real time. Optical flow is also sensitive to the image's noise.

1.6 Item Tracking Video Analytics Software

- ❑ After detecting moving items, Video Analytics track their movement over the video images. Each task requires locating an item tracked from one image to another. This can be done in 2D with a single camera or in 3D combining two views with a known geometric relationship.
- ❑ Many tracking software are based on mathematical methods that make it possible to predict an item's position on a video frame based on its movement in the previous video frames. Tracking several items at the same time poses many challenges. Each item detected in a video frame must be associated with its corresponding item in the subsequent frame. This matching is done based on the items' outlines, their signatures (e.g., corners, area, ratios), or their model of appearance
- ❑ Obstructions (regions hidden by others) represent a key difficulty for tracking items. An Intelligent Video Surveillance system may lose track of an item if it is totally or partially obstructed over a certain period of time. It may also be difficult to separate two items when they are very close or when one hides another.

1.7 Object Sorting and ID

Items detected by a Video Analytics system are usually classified into different categories: human, vehicle, animal. This sorting may be done prior to tracking in order to retain only the trajectories of items that are relevant for surveillance purposes.

In many cases, systems recognize the nature of an entity detected based on its shape attributes and movement properties. For example, a human is usually presented as a form that is taller than being wider. Though a vehicle would be wider than being taller, human posture has clear features, in particular a certain periodicity.

1.8 Item Identification and Recognition

Item identification challenges item recognition further. After finding the class to which an item belongs, it must be identified. With surveillance, and in particular for access control or when searching for a suspect, the goal is to recognize an individual or decipher a vehicle license plate. Vast research and development efforts have been invested in recent years in these two specialized applications.

Recognizing human faces and the pattern of movement of the limbs are the two main biometric tools to identify people in video surveillance. Analysis of the pattern of movement of the limbs provides clues for the preliminary identification of an individual image from a distance in a wide field.

1.9 Multi-Camera Intelligent Video Surveillance Systems

Most intelligent video surveillance systems in particular on IP networks, may be comprised of hundreds of video cameras. Certain times, several video cameras cover the same area. Certain of them are power-operated and can be controlled to capture further details on an event detected in a wide field. For example, the camera can zoom in on an individual penetrating an area in order to identify a person. In certain networks, video cameras are intelligent, i.e., they have their own processing unit. They can exchange data with a central system or directly amongst themselves.

These surveillance camera networks make it possible to follow items over extensive areas. Furthermore, multiple views may help to solve item occlusion complications. In a distributed architecture, analytical processing can be done in parallel, thereby accelerating the analysis and saving on bandwidth by transmitting only metadata.

However, these networks also raise clear complications that the scientific world is trying to solve:

Camera calibration: A procedure that consists of establishing a correspondence between the global reference mark of the scene observed and the camera coordinate system as well as in determining the camera's intrinsic parameters, for example image distortion. This precision process may be tedious for a network with many video cameras. It is preferable to develop software that does not require any camera calibration or self-calibrating methods.

Movement detection: For a power-operated camera, the camera's movement creates an apparent change in the image. Movement detection methods must differentiate between camera movement and independent changes.

Item tracking on several video cameras: In order to be able to track an item from one camera to the next, correspondence must be established between the different views in a common reference point. Camera resetting consists of calculating parameters for transforming the image from one camera to the next based on the change in reference point and the movement model. This procedure uses a prior knowledge of the scene's topology. It allows for an increased 2D or 3D view to be obtained of the scene. Changes in item appearance and positioning over time as well as changes in illumination complicate the resetting process.

Detection of camera tampering: The more video cameras a network has, the more difficult it is to control how they function. These systems must have built-in tools for automatic detection of camera breakdowns and alterations in order to remain functional.

Multi-camera Intelligent Video Surveillance systems are used above all for monitoring extensive or distributed areas, for example transportation systems, banking infrastructures, government institutions, military bases, prisons, strategic infrastructures, centers, hospitals, public buildings, shopping malls and parking lots.

1.10 Video Content Analysis

- Stand-alone and CCTV networks are present in safe cities, transportation, law enforcement and in city first responders infrastructures.
- Video Content Analysis (VCA) systems are intended to improve safety and security in safe cities. It is an open secret that it is increasingly difficult for operators to identify a threat while viewing a bank of monitors; operators miss over 90% of such events. VCA systems assist the operators in focusing on abnormal events.
- **Note**: Video Content Analysis technology should not be confused with legacy video motion detection (VMD), a technology (which is a sub-sector of VCA) that has been in the market for over 20 years. VMD uses simple rules and assumes that any pixel change in the scene is important. One limitation of VMD is that there is an inordinate amount of false alarms.

- Generic CCTV systems are unable to analyze the massive information they collect. Therefore, the detection and interpretation of operations and decision-making in real-time as well as the need to investigate events in retrospect are normally done by human operators. Human response to video image perception suffers from many flaws:
  - Slow response time – poor concentration and vigilance.
  - Need to address information presented on many monitors,
  - Due to congestion of static and dynamic information, it is difficult to find an obvious risk in the scene without defining what exactly to look for.
  - A security officer needs time to study the situation.
  - Human readers have a difficulty in sorting and filtering information.
  - The need for personnel to address info on each and every monitor.
  - Effective supervision requires one person per monitor. In practice, the typical ratio is 10 monitors for 100 cameras and one man in front of 10 monitors.
  - Cost of labor.

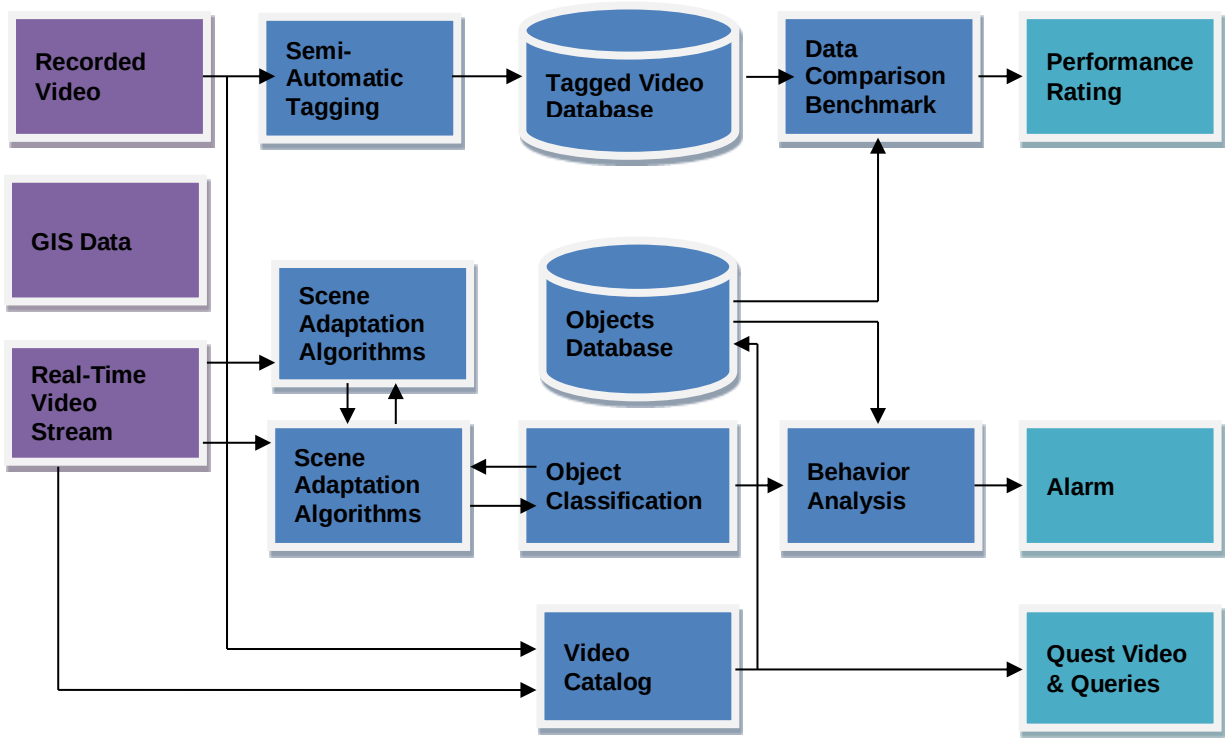
VCA shortens and improves threat detection. When a VCA algorithm detects threat information, it immediately presents the data and image to security personnel.

VCA software provides technological goals, including:

- Focusing the operator on areas of interest
- Analyzing exceptions based on statistical analysis of all the events on video.
- Automatic VCA learning features diverse environmental scenarios and adaptations including:
- Ability to search the video - "Google" video.
- Create a scale for VCA to technologically integrate standardization issues.
- Technological gaps
- Tracking objects – in a dense environment, changing lighting conditions, difficulties of signal / noise.
- Classification of objects in different departments of environmental conditions.
- Adapting to different conditions of a scene - automatic learning routine in the monitored area.

- Identify anomalies automatically in real-time conditions as opposed to sterile conditions.
- Search the object on site under different scenes.
- Tagging for Benchmarking – non-automated or semi-automated tools to examine the component video. Manual examination blocks development of standard criteria to evaluate VCA technologies.

Figure 2 - Video Content Analysis Block Diagram and Data Flow



The use of visual search enhances the value of the captured media. Video analytics technology can greatly reduce the number of operators reviewing useless content and concentrate resources to review activity of potential interest. The technology can be used on both pre-recorded images and real-time video surveillance feeds. Some possible uses of the video analytics software in security and surveillance markets are:

- **Real-time analysis of video surveillance feeds:**
  - Video security/surveillance systems can benefit from the application of image analysis software. The video analytics software can be used to carry out a "continuous" search of a live video feed. An operator could formulate one or more standing queries and the live video can be analyzed in real-time. When the query item(s) are detected in the field of view of the camera, triggers can be activated (send an e-mail, start recording, sound an alarm, etc.)

- In another example, the feed from an altitude camera triggers an event when two or more vehicles of a certain type come to an intersection within a predefined time interval. The software effectively handles this situation even when there is a lot of activity in the video. For instance, even if there were a tremendous amount of pedestrian traffic, the software would not trigger an event unless it identifies an object that physically matches the search criteria.

❑ **Post processing of images/video to find events of interest:**

- Scene analysis is the most common way to segment a video. With today's existing techniques, this segmentation is done at the time of ingestion and is static (i.e., only done once). With video analytics, it is segmented at the time the user carries out the search with the segmentation based on finding objects of interest.
- One possible area this could be applied to is in analyzing surveillance images/video. With dynamic segmentation, it would be very easy to go back and analyze historical footage to find objects that might have recently become more relevant.

An effective Intelligent Video Surveillance system involves detection of one or more biometric signatures in a noisy environment at a distance which minimizes the injury to security personnel and assets. In most cases, the background will be dynamic, differing greatly from that found in a laboratory environment.

Detection methods generate both (false) alarms in the absence of threat and true alarms in the presence of a real threat.

How quickly and at what distance an Intelligent Video Surveillance system is operable is a critical factor.

Remote suspects' detection involves a series of steps:

- Receiving a signature,
- Processing the signature,
- Assessing the results, and
- Determining if a threat is present or not.

Other factors that can impact Intelligent Video Surveillance performance include ambient environmental conditions and other people present in the environment.
Intelligent Video Surveillance System Performance: System performance is a function of sensitivity and specificity of the deployed technologies. When orthogonal detection systems are used, both false positives and false negatives will occur. Obviously, the specific capabilities of the component technologies as well as how well the technologies are integrated will impact system performance.

1.10.1 Automated Analysis of Video Surveillance Data

In a typical system for event detection, a set of processing stages need to be performed. First, the relevant item present in the observed scene needs to be detected. In the next step, item classification should be performed which allows differentiating various item types. Such a procedure allows for further analysis depending on the item category. Further processing can lead to event detection which could be associated to certain threat related behavior. This data could be useful for the system operator in improving his awareness about security compromising situations. Many more processing building blocks could be a part of such a system which could simplify and improve operators 'work, such as face recognition. To build a feasible smart surveillance system, currently much emphasis is on making robust video processing software. Hence, a large variety of approaches for acquiring similar effects can be found in the literature. The following sections introduce typical Intelligent Video Surveillance processing building blocks. A general description of their purpose and known methods for accomplishing certain goals are presented.

1.10.2 Item Detection

In many cases, item detection is usually the first building block in the video processing pipeline in smart surveillance systems. This building block is designed to detect the relevant items which are observed in the scene. This goal can be achieved in various ways.

One of the groups of methods employs so-called optical flow software which depending on clear algorithm, analyses the scene for characteristic points or estimates the motion on the basis of gradient calculated for the whole image. However, this set of methods is in general quite computationally expensive and is not suitable for real time processing of high resolution images and high frame rate videos. Still, certain improvements and simplifications to the optical flow calculation algorithm can be found in the literature, thereby making it useful for clear tasks. This approach is generally a good choice in two cases. First, when a video stream from a non-static camera needs to be analyzed. In such situations, optical flow is very suitable as it allows for compensating the global motion detecting certain local changes. The second case is related to crowded scenes where the data about single items and their movement is difficult to acquire. By using image flow based methods such as a general crowd, motion trend can be estimated.

In the next group of methods, before item detection occurs, a static image (called background model) representing the stationary parts of the observed scene is built. To extract the dynamic parts of the processed frame, the current image and the background model are compared. The regions which differ much from the background model are marked as moving items and considered as the foreground. This type of technique is usually referred to as background

subtraction. The simplest way to acquire an adaptive background model is to utilize frame-averaging algorithm where the model is a cumulative average of consecutive video frames with a set learning factor.

Unfortunately, the adaptation offered by such algorithms is insufficient for use, especially for typical outdoor conditions. Therefore, methods that are more sophisticated have been proposed, among which Gaussian Mixture Model based on background subtraction is considered to be one of the most popular. This approach allows for more robust item detection because of higher adaptability and therefore lower sensitivity to global light changes. This technique introduces only temporal relations in the procedure of building a background model, but a number of software which involve spatio and spatio-temporal relations are proposed in the literature.

Another set of methods detect items on the basis of their known appearance. This approach is useful in cases where it is not feasible to generate a background image, such as with a moving camera platform or when the background is predominantly obscured by the foreground image. This set of methods requires a single image or a set of item representations to be detected. On the basis of certain features extracted from the supplied models, the desired item or a part of that item can be detected. This approach can be useful in cases of face or registration plate detection. The most popular algorithm among this group utilizes Haar-like features (Haar-like features are digital image features used in object recognition. They owe their name to their intuitive similarity with Haar wavelets and were used in the first real-time face detector and a set of weak cascade classifiers trained on a provided dataset.

The contents of the dataset depend on the problem, however a number of already trained classifiers or at least prepared models can be found on the internet. If the item color spectra of the detected item are known, a mean-shift based approach can be applied. This algorithm performing image segmentation and its extension called CAMShift are suitable for real time processing. Like most of the advanced methods, it introduces an adaptation procedure and therefore can be quite robust.

1.10.3 Background Subtraction: Gaussian Mixture Based Software

Detection of moving items is usually the first stage of a video processing chain and its results are used by further processing building blocks. Most video segmentation software usually employs spatial and/or temporal data in order to generate binary masks of items. However, simple time averaging of video frames is insufficient for a surveillance system because of limited adapting capabilities. The solution applied in the framework employs spatial segmentation for detection of moving items in video sequences using background subtraction.

This approach is based on modeling pixels as mixtures of Gaussians and using an on-line approximation to update the model. This technique proved to be useful

in many applications as it is able to cope with illumination changes and to adapt to the background model accordingly to the changes in the scene, such as when motionless foreground items eventually become a part of the background. Furthermore, the background model can be multi-modal, allowing regular changes in the pixel color. This makes it possible to model such events as trees swinging in the wind or traffic light sequences.

Background modeling is used to model current background of the scene and to differentiate foreground pixels of moving items from the background.

Item segmentation is supplemented with shadow detection and a removal module. The shadow of a moving item moves together with the item and as such is detected as a part of the foreground item by a background removal algorithm. The shadow detection technique is based on a concept that while the color component of a shadowed background part is generally unchanged, its brightness is significantly lower. Every pixel recognized as a part of a foreground item during the background subtraction process is checked whether it belongs to a moving shadow. If the current pixel is darker than the distribution, it is assumed to be a shadow and is considered as a part of the scene background.

Another approach shows the shadow‘s shape, size, orientation, luminosity, originating position and appearance model exploited to determine the color distributions of both the foreground and shadow classes. This is achieved by skeletonization and spatial filtering process which is developed for identifying components in the foreground segmentation that are most-likely to belong to each class of feature. A pixel classification mechanism is then obtained by approximating both classes of feature data by Gaussian parametric models. This work is further extended in where novel k-nearest neighbor pixel classifier is proposed, which is applied on pixels previously classed as foreground during detection process in real time.

As a result of background modeling, a binary mask denoting pixels recognized as belonging to foreground items in the current frame is obtained. It needs to be refined by means of morphological processing in order to allow item segmentation. This process includes finding connected components, removing items that are too small, morphological closing and filling holes in regions. Additionally, an algorithm for shadow removing from the mask using morphological reconstruction is applied. The morphological reconstruction procedure involves two binary images: a mask and a marker. In the mask image, all pixels belonging to either the moving item or the shadow have a value of one, and all the background pixels have zero value. The marker is obtained by applying an aggressive shadow removal procedure to the item detection result, so that all the shadow pixels are removed.

Another challenge for item segmentation is related to detection and removal of ghosts caused by the starting or stopping of items. Ghosts mainly appear in two cases: In the first case, when a moving item becomes stationary, it will be adapted (merged) into the background, and when it starts to move again later,

there will be a ghost left behind. In the second case, an existing item that belongs to the background starts to move (such as parked vehicle) and will also cause a ghost to be formed. To tackle this problem, a comparison of the similarity between the edges of the detected foreground items and those of the current frame based on item level knowledge of moving items is done.

1.10.4 Background Subtraction

When two stereo video cameras are available, one model can be used to create two background models, i.e. a color intensity background model and a stereo background model. The combination of both models allows performing more robust and effective segmentation even in very crowded backgrounds.

The main disadvantage of this model is that it does not take into consideration the correlation between neighboring pixels. In the novel background estimation technique, the co- variation of grey levels within the incoming images using principal component analysis to generate the eigen-backgrounds is proposed. Rather than accumulating the necessarily enormous training set, this technique builds and adapts the eigen-model online. The number of significant modes as well as the mean and covariance of the model is continuously adapted to match the background conditions.

Consequently, for each incoming image, a reference frame is hypothesized efficiently to perform background segmentation from a subsample of the incoming pixels. In turn, it proposes to take advantage of periodic variation of background appearance over time which is detected in the temporal frequency domain. This analysis allows estimating the period for each pixel, and pixels with the same periodicity are grouped into regions. A Markov model can then be constructed in which each state models the first and second order statistics of the appearance at a given phase of the period. The state values are updated online. As a consequence, the foreground can be segmented accurately from this time-varying background.

1.10.5 Item Detection Using a Single-image Software

In certain scenarios, the estimation of the background model may be problematic due to a highly dynamic and complex background. However, when training data is available, the item detection can be performed without the need of creating a background model.

In recent years, several planners have been proposed to detect categories of items from a single image. The categories of items not only include faces and people, but also extend to less common sets. When there are multiple categories of item in the same system, it is usually described as an item recognition system (rather than as an item detection system). The PASCAL Visual Item Challenge includes examples of a bicycle, boat, bus, car, motorbike and train: these are

naturally useful in a surveillance context. For instance, the item detection can be achieved by building the cluster boosted tree structure for a multi-view classification based on edgelet features or using predefined D models which are combined with local patches of histograms of oriented gradients.

1.10.6 Item Tracking Software

Item tracking software usually follows the item detection. This building block allows introduction of a relation between items detected in consecutive video frames.

Tracking of human motion is a very active research field which has been addressed in various ways. In the case of a single item tracking, it uses a limb tracking system based on a 3D articulated model and a double tracking strategy. Its key contribution is that the 3D model is only constrained by biomechanical knowledge about human bipedal motion, instead of being tailored to a clear activity or camera view. This double tracking strategy is based on a Kalman digital filter for global position tracking and a set of particle filters for body parts tracking. Since a space of human motion is often high dimensional, tracking can be performed efficiently in a low dimensional space. This is achieved by the advanced limb correction building block which proposes graph-based particle digital filter to deal also with stylistic variations of motion.

To track multiple items in complex backgrounds, the bank of Kalman filters is exploited to track each subject independently in a scene. In order to facilitate data association and track management, a color model is created for each person. Tracking of multiple items can be enhanced by integration data of self-calibrated ground planes or depth probability density functions. In turn, it presents a real-time algorithm which is based on blob matching software.

First, foreground pixels are detected using luminance contrast and grouped into blobs. Then, blobs from two consecutive video frames are matched to create matching matrices. Tracking is performed using direct and inverse matching matrices. This technique successfully solves blobs merging and splitting during tracking. Finally, it combines several advanced software, for example adaptive intensity-plus-color mixtures of Gaussians, region based representations, Kalman filters, scene model and a Bayesian network to perform an online multi-item tracking in realistic situations from a single fixed camera.

To deal with large obstructions during tracking, a constant acceleration motion model is used to track items where three predictors are employed simultaneously along with a least square correlation stage to select the most likely item position.

1.10.7 Kalman Filtering Techniques, Region Segmentation

Another challenging problem is tracking across multi-camera networks. The proposed system comprises of two processing stages, operating on data from first a single camera and then multiple video cameras. The single-view processing includes change detection against an adaptive background and image-plane tracking to improve the reliability of measurements of occluded players. The multi-view process uses Kalman trackers to model the player position and velocity to which the multiple measurements input from the single-view stage are associated. Multiple video cameras with overlapping views can also help with resolving issues of item obstructions. For instance, tracking is performed in D using the Kalman digital filter together with the ground plane homographic constraint.

1.10.8 Kalman Filters Application to Track Moving Items

After the moving items are found in each consecutive camera frame, movement of each item on a frame-by frame basis is needed. This is the task of an item tracking building block. For each advanced detected moving item, a structure called a tracker is created. The position of the item in the current camera frame is found by comparing the results of item detection (the blobs extracted from the image) with the predicted position of each tracker. The prediction process estimates the state of each tracker from the analysis of the past tracker states. An approach based on Kalman filtering is used for prediction of trackers' state.

A relation between a tracker and a blob is established if the bounding box of the tracker covers the bounding box of an item by at least one pixel. There are certain basic types of relations possible; each of them requires different actions to be taken. If a certain blob is not associated with any tracker, an advanced tracker (Kalman filter) is created and initialized in compliance with this blob. If a certain tracker has no relation to any of the blobs, then the phase of measurement update is not carried out in the current frame. If the tracker fails to relate to a proper blob within several subsequent video frames, it is deleted. The predictive nature of trackers assures that moving items whose detection through background subtraction is temporarily impossible are not lost (such as when a person passes behind an opaque barrier).

If there is an unambiguous one-to-one relation between one blob and one tracker, this tracker is updated with the results of the related blob measurements. However, if there is more than one matching blob and/or tracker, a tracking conflict occurs. The authors proposed the following algorithm for conflict resolving. First, groups of matching trackers and blobs are formed. Each group contains all the blobs that match at least one tracker in the group and all the trackers that match at least one blob in the group. Next, all the groups are processed one by one. Within a single group, all the trackers are processed successively. If more than one blob is assigned to a single tracker, this tracker is

updated with all blobs assigned to it merged into a single blob. This is necessary in case of partially covered items (such as a person behind a post) that cause the blob to be split into parts.

In other cases, all the matching blobs are merged and the tracker is updated using its estimated position inside this blob group. This approach utilizes the ability of Kalman trackers to predict the state of the tracked item, provided it does not rapidly change its direction and velocity of movement and also that the predicted state of the Kalman digital filter may be used for resolving short-term tracking conflicts. The estimated position is used for updating the tracker position; change of position is calculated using the predicted and the previous states. The predicted values of size and change in size are discarded and replaced by values from the previous digital filter state in order to prevent disappearance or extensive growth of the tracker if its size was unstable before entering the conflict situation. Therefore, it is assumed that the size of the item does not change during the conflict.

1.10.9 Partially Observable Markov Decision Process, Intelligent Video Surveillance Systems

A Partially Observable Markov Decision Process (POMDP) is a generalization of a Markov Decision Process. A POMDP models an agent decision process in which it is assumed that the system dynamics are determined by an MDP, but the agent cannot directly observe the underlying state. Instead, it must maintain a probability distribution over a set of possible states based on a set of observations and observation probabilities and the underlying MDP.

The POMDP framework is general enough to model a variety of real-world sequential decision processes. Applications include robot navigation problems, machine maintenance, and planning under uncertainty in general. This framework originated in the Operations Research community and was later taken over by the Artificial Intelligence and Automated Planning communities.

An exact solution to a POMDP yields the optimal action for each possible belief over the world states. The optimal action maximizes (or minimizes) the expected reward (or cost) of the agent over a possibly infinite horizon. The sequence of optimal actions is known as the optimal policy of the agent for interacting with its environment.

A system being developed by Amato et al, at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) can perform Intelligent Video Surveillance analysis more accurately and in a fraction of the time it would take a human CCTV camera operator.

The system uses POMDP to reach a compromise on accuracy so that it does not trigger an alarm every time a cat walks in front of the camera, for example the speed needed to allow security staff to act on an event as quickly as possible.

For Intelligent Video Surveillance systems, users have typically a range of different computer vision algorithms they could use to analyze the video feed. These include skin detection algorithms that can identify a person or an item in an image, and background detection systems that detect unusual objects or when something is moving through the scene.

To decide which of these algorithms to use in a given situation, the MIT system first carries out a learning phase in which it assesses how each piece of software works in the type of setting in which it is being applied such as an airport. To do this, it runs each of the algorithms on the scene to determine how long it takes to perform an analysis and how certain it is of the answer it comes up with. It then adds this information to its mathematical framework known as a partially observable Markov decision process (POMDP).

For any given situation, if the system wants to know if an intruder or object has entered the scene, it can decide which of the available algorithms to run on the image and in which sequence to give it the most information in the least amount of time. According to Amato, the lead system developer, "We plug all of the things we have learned into the POMDP framework and it comes up with a policy that might tell you to start out with a skin analysis for example, and then depending on what you find out, you might run an analysis to try to figure out who the person is or use a tracking system to figure out where they are (in each frame), and you continue doing this until the framework tells you to stop essentially when it is confident enough in its analysis to say there is a known terrorist here for example, or that nothing is going on at all."

Like a human detective, the system can also take context into account when analyzing a set of images, Amato said. So for instance, if the system is being used at an airport, it could be programmed to identify and track particular people of interest and to recognize objects that are strange or in unusual locations, he said. It could also be programmed to sound an alarm whenever there are any objects or people in the scene, when there are too many objects, or if the objects are moving in ways that give cause for concern.

In addition to port and airport security, the system could monitor video information obtained by a fleet of unmanned aircrafts, Amato said. He also said it could be used to analyze data from weather monitoring sensors to determine where tornados are likely to appear or information from water samples taken by autonomous underwater vehicles. The system would determine how to obtain the information it needed in the least amount of time and with the least possible number of sensors.

According to Matthijs Spaan, an artificial intelligence researcher at Delft University of Technology in the Netherlands, the work demonstrates how artificial intelligence decision-making techniques can benefit data-intensive applications such as automated video surveillance. "Video processing has high computational demands and the work shows how POMDPs can be applied to dynamically trade off computation cost with prediction accuracy," he says. "The POMDP model

excels at decision-making regarding uncertain information, in this case whether an intruder is present or not."

1.10.10 "Splitting" Items Algorithms

A special case of tracking conflict is related to splitting items, such as if a person leaves a luggage and walks away. In this situation, the tracker has to follow the person and an advanced tracker needs to be created for the luggage. This case is handled as follows. Within each group of matching trackers and blobs, subgroups of blobs separated by a distance larger than the threshold value are found. If there is more than one such subgroup, it is necessary to split the tracker: select one subgroup and assign the tracker to it and then create an advanced tracker for the remaining subgroup. In order to find the subgroup that matches the tracker, the image of the item stored in the tracker is compared with the image of each blob using three measures: color similarity, texture similarity and coverage. The descriptors of the blob are calculated using the current image frame. The descriptors of the tracker are calculated during tracker creation and updated each time the tracker is assigned to only one blob (no conflict in tracking).

After the conflict resolving is done, the tracking procedure finishes with creating advanced trackers for unassigned blobs and removing trackers to which no blobs have been assigned for a defined number of video frames. The process is repeated for each camera frame, allowing for tracking the movement of each item.

The figure below is an example of item tracking with conflict resolving, using the procedure described here. A person passes by a group of four persons walking together. During the conflict, positions of both items are estimated using the Kalman digital filter prediction results. When these two items become separated again, assignment of trackers to blobs is verified using the color, texture and coverage measures. As a result, both items are tracked correctly before, during and after the conflict occurs.

An essential part of a smart surveillance system is the items classification building block. Differentiating between items allows for more detailed analysis of item behaviors. If the items present in the analyzed material can be assigned to a specified class, further detected actions or other applied procedures could be done depending on the type assigned to the item.

Item classification in Intelligent Video Surveillance systems is not an easy task mainly due to variation of camera viewpoints. Several methods have been proposed to solve this problem which can be divided into two general groups: feature-based and motion-based. Feature based approach utilizes the data on item spatial signatures commonly related to shape or texture descriptors. One of the clear methods uses the item's real dimension to decide about its type. This

approach considers possible perspective differences between video cameras but requires a calibration procedure to be applied.

Other approaches exploit more advanced properties acquired from the detected items. Although in this case no calibration is required, a large set of prepared item models are needed for proper classifier training. For the purpose of model and item parameterization, a set of various features can be used. The most common parameters present in the literature include SIFT descriptor and a number of contour and region based shape descriptors. Many other description methods such as wavelet coefficients can be found in the literature. A particular technique based on the item's shape and its contour description is presented deeper in the following subsections

Item categorization is performed by the built classifier on the basis of the calculated feature vector. Again, a number of different approaches for the classification process exist. Feature-based methods applying various distance metrics as well as machine learning software like Support Vector Machines (SVM) and Artificial Neural Networks (ANN) can be found in the literature.

Another approach found in the literature employs Motion History Images and Recurrent Motion Images as a technique not only for action recognition but for item classification processes as well. These methods are based on the property that certain items like cars are more rigid in comparison to other items like people. Therefore, by analyzing the spatio-temporal signatures, it is possible to differentiate between these items.

1.10.11 Dimension Based Items Classifiers

Dimension based item classifier is an example of a classifier which utilizes a set of thresholds for the purpose of assigning items to a proper category, i.e. Human or Vehicle. Due to the perspective distortion in a typical camera image, it is not possible to use the size of an item measured in image pixels for classification purposes, but the width and the height of the item in physical units (i.e. meters or inches) has to be known. Therefore, a conversion between the camera coordinates and the physical coordinates has to be defined and this can be achieved by means of the camera calibration procedure. Various calibration methods were proposed, for example in one method the calibration technique requires marking several points in the area observed by the camera and measuring their positions relative to each other, both in the real world and in the computer image. One of these points is usually set as an origin of the world coordinate system. The calibration algorithm processes pairs of the world coordinates and the image coordinates of each point in order to calculate conversion coefficients, describing translation and rotation of the camera relative to the world coordinates, camera lens distortions, and camera focal length.

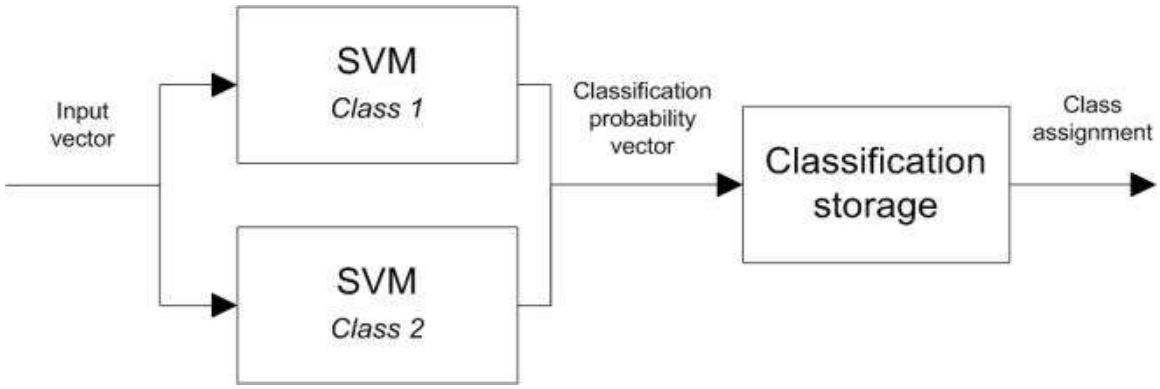
1.10.12 Shape Based Item Classifiers

Intelligent Video Surveillance cameras can be placed in various ways: on top of the building, attached to a pole, or hung under the ceiling. This implies that the same item is observed from different cameras and can be represented by a different silhouette. Moreover, the item itself can rotate around its own axis creating more silhouette variations. This shows the complexity of the considered problem.

Shape based item classification has a pattern recognition problem and requires a set of items for a proper classifier training.

For the purpose of classification, Support Vector Machines were chosen. Its goal is to find a hyper plane which divides two sets of data with the biggest margin between them. The SVM classifier used utilizes the RBF which is said to be the first and best choice. Optimal SVM cost and gamma parameters are set during training using the grid-search technique explained in the literature. For the training and classification process, One vs. All algorithm is used. Therefore in case of classes, two binary classifiers are built. To obtain classification probabilities, a statistical model is built using the algorithm applied in Libsvm (Library for Support Vector Machines) library.

Figure 3 - Classification Process Scheme



1.10.13 Event Detection Methods

Image-processing software is used to automatically detect incidents and take measurements, thereby relieving the control room personnel of the effort in finding out where events take place. Events of particular interest may include abnormal stationarity, queuing, intrusion detection, loitering, unattended luggage detection as well as closely related complications, for example action recognition. Certain events could be detected on the basis of a number of user predefined rules while others required more complex methods involving behavior analysis.

In general, event detection approaches can be divided into two groups:

The first case can be presented as an unsupervised system which builds a kind of probabilistic behavior pattern. This way, the anomalies which do not fit the estimated model can be detected. The main advantage of this approach is that no user input regarding the events is required and that it adapts dynamically to the changing conditions. At the same time, it offers a low control of the detected events as they are not strictly defined. Therefore, a number of non-relevant situations can be detected if certain situations are rare in a general context.

The second more frequent case is when a rule based approach is applied. Here, a number of atomic events like: item stopped, item left the scene, and item entered area are defined. On the basis of these elementary events, complex rules can be built that allow for clear situation detection.

In respect of the analyzed item, the event detection problem can concern items in general or people in particular. In the second case, a single person, groups or crowds need to be considered. Various events are possible to be detected in each of these cases. For a crowd, in many cases, its flow-based clear activity and certain anomalies from the main trend can be detected. As for single people, primitives-based event detection as well as more complex behavior recognition can be applied.

An important problem with detection of left and removed items is that due to the nature of the background subtraction algorithm, leaving an item in the scene causes the same effect as removing an item that was a part of the background (such as a luggage that remained stationary for a prolonged time).

1.10.14 Vision-based Human Action Recognition

Vision-based human action recognition is a high level process of image sequence analysis. Any robust action recognition system should be able to generalize over variations of style, view and speed within one class and differentiate between actions of different classes. This is usually achieved by learning so called action models from pre-acquired training datasets. Subsequently, these action representations are used for classification of unseen action instances.

The popular approach for modeling action is referred to as Bag of Words which models an action as a large visual vocabulary (dictionary, codebook) of discriminative code words. This visual dictionary is formed by the vector quantization of local feature descriptors extracted from images using for instance the k-means algorithm. A sequence of images is summarized by the distribution of code words from the fixed codebook by computing a histogram of code word occurrences based on the assignment of local descriptors. Action classification is performed by constructing a feature vector for videos to be related based on the defined dictionary.

Action video sequences are high dimensional because of the human motion complexity. However, different instances of the given action reside only in a part of the entire feature space. This subspace can be considered as a nonlinear manifold embedded in a space of video frames. As a result, the discriminative and low dimensional manifold of the action can be discovered by a dimensionality reduction process. For instance, a view dependent action recognition can be performed by applying temporal Lapacian Eigenmaps on training videos to extract intrinsic characteristic of the action followed by a nearest neighbor classification schema in the obtained low dimensional space . This concept is further extended in and to perform view independent action recognition. Alternatively, the temporal extent of action can be represented using Hidden Markov Model and the low dimensional Self Organizing Map, so the subsequent inference of action label can be performed in a probabilistic manner.

1.10.15 3D Derived Egomotion

Egomotion is defined as the 3D motion of a camera within an environment. In the field of computer vision, egomotion refers to estimating a camera's motion relative to a rigid scene. An example of egomotion estimation would be estimating a car's moving position relative to lines on the road or street signs as observed from the car itself. The estimation of egomotion is important in autonomous robot navigation applications.

The goal of estimating the egomotion of a camera is to determine the 3D motion of that camera within the environment using a sequence of images taken by the camera. The process of estimating a camera's motion within an environment involves the use of visual odometry techniques on a sequence of images captured by the moving camera. This is typically done using feature detection to construct an optical flow from two image frames in a sequence generated from either single cameras or stereo cameras. Using stereo image pairs for each frame helps reduce error and provides additional depth and scale information.

Features are detected in the first frame and then matched in the second frame. This information is then used to make the optical flow field for the detected features in those two images. The optical flow field illustrates how features diverge from a single point called the focus of expansion. The focus of expansion can be detected from the optical flow field indicating the direction of the motion of the camera and thus providing an estimate of the camera motion.

There are other methods of extracting egomotion information from images as well, including a method that avoids feature detection and optical flow fields and directly uses the image intensities.

1.10.16 Path Reconstruction Software

Path reconstruction is a method of particular interest. Several approaches to this issue have already been proposed in the past. They depend on the camera arrangement within the multi camera system. A number of complications need to be tackled for a robust route reconstruction. For instance, spatio-temporal camera relations need to be discovered and differences in images from various video cameras should be taken into account for the purpose of proper item description.

In the following subsections, a number of approaches for item tracking in multi camera systems found in the literature is briefly presented. After this description, a clear and already applied solution introducing the basic idea of route reconstruction is shown with more details.

One technique is based on given spatio-temporal topology of the monitoring network. The whole system is described as a graph which is used to search and re-recognize the same item, in particular video cameras (as the item is moving). The most important thing in this method is the item recognition algorithm because it is crucial for the effectiveness of the described method. The proposed network model assumes two categories for item state: hidden and visible video cameras. The features of the item are stored as parameter vectors and their similarity is the main input for re-recognition algorithm. Two phases are suggested in the proposed method:

Modified particle digital filter modeling of hidden (unobserved by cameras) places, where items are not seen. In any FOV particle, a digital filter is chosen (instead of Kalman filter) because the process of tracking many items is non-linear and hence a lot of noise is expected.

1.10.17 Video Cameras Spatial Gap Mitigation Software

This technique uses a great amount of video data and displays "activity maps" using temporal correlation between pairs of video cameras and similarity of particular items. Loci of entry and exit points in video cameras field of view are detected and used to recognize entry and exit events. In places where the camera network is "blind", virtual states are added. Conditional probabilities define rate of transitions between particular pair of states and on this basis, the topology of the network is discovered. The obtained topology is a spatio-temporal one. The entire technique can operate unsupervised and doesn't need any calibration of the video cameras. This technique can generate redundant connection between cameras.

Another technique of tracking items between non-overlapped FOVs of cameras uses two types of data acquired from the monitoring system. The first type is appearance of the observed item and the second is spatio-temporal data. Two-dimensional histograms are used: histograms of color for whole item (1st

dimension) and histograms of color for a part of the item, for example: head, torso and legs (2nd dimension). The spatio-temporal data are acquired on the basis of transition time distribution. This distribution function is modeled as a mixture of Gaussian distributions.

In this technique, three types of Gaussian distribution are chosen and each distribution describes the way of walking. Certain people walk slowly, while other people very quickly. Each of these Gaussian distributions is weighted in order to fit the mixture of Gaussian distributions to the histogram of transition time between particular pairs of cameras. Weights in the proposed model are estimated by "expectation maximization" algorithm. Experiments performed by authors of this paper validate correctness of used methods.

1.10.18 Networked Cameras Tag and Track Software

This algorithm is designed for tracking an item across multiple non-overlapped cameras in real time. It is based on a probabilistic framework which integrates data from a variable number of heterogeneous building blocks in real time. These informative clues include:

- Data on a target (such as appearance and position) obtained from tracking building blocks, working on single and calibrated video cameras
- Data from people re-identification building blocks, working across video cameras
- Prior knowledge about a scene, which originates from the camera network topology

This process starts with the initialization, i.e., a manual selection by the operator to follow this chosen target automatically across different non-overlapped CCTV cameras. This is achieved by taking advantage of several technologies to tackle challenging complications of static and dynamic obstructions, crowded scenes, poor video quality and real-time procedure.

First, foreground detection exploits the presence of many periodically-changing background elements in indoor scenes (escalators, scrolling advertisements, flashing lights). In many cases, it detects and models pixels exhibiting periodic changes in color and uses this data to predict the pixel color in subsequent video frames in order to improve on foreground detection. Then the extracted moving foreground is fed into the Kalman Digital filter to perform a single camera tracking. For the problem of data association between video cameras in a network, a novel color correction technique is proposed to allow a robust appearance comparison between targets. It can learn differences in camera color responses up to second-order statistics; the algorithm is completely unsupervised and can automatically detect whether it has gathered enough data for a reliable color correction.

In turn, the camera network layout is learnt automatically based on an activity-based semantic scene model. In the first step, regions of interests, for example entry/exit zones, junctions, paths, and stop zones, are determined from motion tracks. Then the topological relations between them are expressed in probabilistic manner using a Bayesian belief network.

Finally, a probabilistic approach fuses all heterogeneous data coming from the described building blocks. The framework finds a target identity that is globally and most likely to be the one intended by the operator. The solution is re-computed in each frame using only the state of the system in the previous frame, thus avoiding the computational overhead of optimizing a long track and allowing the multi-camera tracker to work in real-time.

1.10.19 Visual Intelligence Technologies

Several research programs (e.g., the US Defense Advanced Research Projects Agency (DARPA) “Mind's Eye”) are aimed to develop a new kind of artificial intelligence (AI) known as “visual intelligence”. It aims to enable machines to robustly recognize and reason about activity in full-motion video footage.

Four performance tasks to encourage development of robust visual intelligence systems are:

- recognition: ‘visual intelligence systems are expected to judge whether one or more verbs are present or absent in a given video
- description: ‘visual intelligence systems are expected to produce one or more sentences describing a short video suitable for human-machine communication
- gap-filling: ‘visual intelligence systems are expected to resolve spatiotemporal gaps in videos, predict what will happen next, or suggest what might have come before
- Anomaly detection: ‘visual intelligence systems are expected to learn what is normal in longer-duration videos and detect anomalous events.

Figure 4 - Visual Intelligence Process

1.10.20 The Visual Intelligence Process

1.10.20.1 Visual Processing

The visual processing of a scene starts with the detection of meaningful objects and their properties. The detection of objects is performed in two ways. First, moving objects are detected by background subtraction. This provides a segmentation of all moving humans, vehicles and other objects in the scene in case they are moving during the activity.

Second, a trained object detector for specific classes like humans and cars is used to detect instances in single frames. This enables the detection of objects without requirement of their motion. Many activities of humans entail for example, small arm movement while the whole body stays in the same position. When objects have been detected, they are tracked through the video stream. The object's position over time is an essential property that directly relates to its action. The tracker tracks the dimensions and colors of detected objects to find new object positions in subsequent frames. The model is updated to follow objects whose appearance changes slowly over time.

Stationary objects can perform actions that are not captured through coarse movement of the whole object. Therefore, a more detailed description of pose and body part movement is computed. Feature descriptors of limb movement based on spatiotemporal interest points, skin color, structural and statistical motion descriptors and salient regions are computed for all objects recognized as humans (by means of the class-specific person detector).

1.10.20.2 Fusion Engine

The purpose of the fusion engine is to filter and possibly fuse the tracked objects in order to form entities. Entities should correspond to the real-world entities like a person, bike or car that contribute to the observed action in the scene. Only the entities – a subset of the detected and tracked objects – are selected for further processing. The output of the fusion engine is a container for each entity which includes the track information and low-level visual features. To limit computation time, there is a delayed execution of several features in visual processing which are only determined where confident entities are found.

1.10.20.3 Event Description

The aim of event description is to raise the level of abstraction from the low-level features towards the object or situation level that is desired to express the rules of an expert system for action recognition. The event properties and derived rules are our way of encoding world-knowledge about the verbs. The properties are related to physical world properties and are based on a taxonomy that positions a

verb in a semantic hierarchy and makes explicit how humans assess and describe events. Three types of event properties are generated; single-entity event properties, entity-pair event properties and global event properties. The first type of events describes properties of one entity (for example “the entity is moving fast”). The second type of events describes the relation between two entities (for example “the distance between two entities is decreasing”). The third type of events describes global properties of the scene (for example “there is only one entity present”).

1.10.20.4 Reasoning

The reasoning component retrieves information on entities and relations present in the video clip from event description. Based on this information, the component infers and describes the behavior of the entities observed in the clip and reports this to the reporting component. In order to do so, the component can be trained on the ground truth available for the observed clips.

The reasoning component contains four independent classifiers:

- RBS: Rule-Based System. A set of manually created rules with several spatio-temporal conditions on event properties is mapped onto a set of verbs. For several rules, it was not possible to define each rule, for example due to missing event properties. A multi-hypothesis partial matcher has been designed, which uses the best match per rule.
- RF-TP: Random-Forest Tag-Propagation. Also a rule-based recognizer, here the rules are learnt from an abundant set of decision trees (i.e. a random forest). The novelty of the usage of these rules is to consider the similarity of distributions over them. The core point of the RF-TP method is that it models the probability of a verb on the current vignette (or video clip) as a consequence of the similarities with all of the previously seen vignettes and the verbs active in those vignettes.
- RTRBM: Recurrent Temporal Restricted Boltzmann Machine. A generative statistical learner that incorporates temporal relations and evidence from observations (in our case event properties).
- HUCRF: Hidden-Unit Conditional Random Field. Similar to RTRBM, yet now a discriminative learner. Both the classifiers (RTRBM, HUCRF) use a condensed version of the event properties which is projected such that it no longer can be distinguished which entity has which property due to implementation constraints.

1.10.20.5 Reporting

This component reports the results provided by the reasoning component in a pre-defined format for each task (i.e. recognition, description, gap-filling and anomaly detection). The output for the recognition task consists of a vector of dozens of probabilities indicating signal strength across the set of verbs for each video clip. The soft probability assignment is converted to a binary presence or absence value of a verb by applying a threshold.

1.11 Video Analytics Challenges

There are several constraints to the applicability of video analytics software which can be summarized as follows. Firstly, such software works on data from a static camera. This allows image differencing software and relatively straightforward calibration'– the definition of the relationship between the camera pixel co-ordinates and the ground or map co-ordinates. This is necessary for global tracking results. If the camera is not static, then both of these software can still be used – but with a more sophisticated approach.

Firstly, image differencing software needs to include techniques for categorizing any observed differences into a) moving items and b) resulting from movements of the camera. Secondly, to work with moving video cameras, calibration methods need to receive data from the telemetry system about the parameters of the movement. This places greater requirements on both the system capabilities and also the automated processing of the geometry to establish the correct calibration. A second constraint on video analytics systems is their capability in crowded scenes. A significant proportion of the software uses a connected components analysis to estimate trajectories etc. As the scene gets increasingly crowded, the components are increasingly inter-connected, rendering this approach less effective.

A third constraint is the available illumination in the scene. With outdoor scenes, the illumination is variable and generally insufficient around night-time. Certain indoor scenes, for example train stations, are traditionally brightly lit, though others (e.g., Railway platforms, rest areas) have lower levels of illumination for economic or cultural reasons.