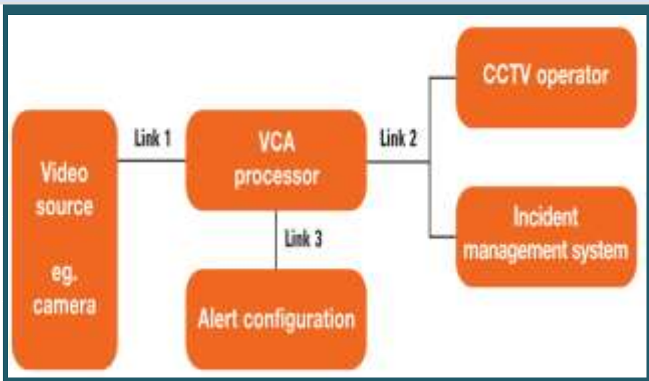
2015

Video Content Analysis (VCA)



Homeland Security Research Corp.

Video Content Analysis (VCA)

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

1. Present and Pipeline Video Analytics (VA) & Video Content Analytics (VCA) Technologies.....4

1.1. Video Content Analysis (VCA) 4

1.1.1. Introduction.....4

1.1.2. VCA Systems 4

1.1.3. Pipeline VCA Research 7

1.1.4. Video Content Analysis Software 8

1.1.5. Present Limitations of Video Analytics Software.....8

1.1.6. Automated Analysis Of Video Surveillance Data 9

1.1.7. Item Detection 9

1.1.8. Background Subtraction: Gaussian Mixture Based Software 10

1.1.9. Background Subtraction 12

1.1.10. Item Detection Based on a Single-Image Algorithm 12

1.1.11. Item Tracking Software.....13

1.1.12. Kalman Filtering Techniques, Region Segmentation.....14

1.1.13. Kalman Filters Application To Track Moving Items 14

1.1.14. Partially Observable Markov Decision Process, IVS Systems.....15

1.1.15. "Splitting" Items Algorithms.....17

1.1.16. Dimension Based Items Classifiers 18

1.1.17. Shape Based Item Classifiers.....18

1.1.18. Event Detection Methods.....19

1.1.19. Vision-Based Human Action Recognition 20

1.1.20. 3D Derived Egomotion.....21

1.1.21. Path Reconstruction Software 22

1.1.22. Video Cameras Gap Mitigation Software.....22

1.1.23. Networked Cameras Tag and Track Software.....23

List of Tables

Table 1 - Video Content Analysis Functions and Description 4

Table 2 - Pipeline VCA: Sub-Systems, Technology, Technology Challenges and Target Performance 7

List of Figures

Figure 1 - VCA Block Diagram and Data Flow..... 6

Figure 2 - Unattended Item Detection Algorithm Flow Chart..... 20

1. Present and Pipeline Video Analytics (VA) & Video Content Analytics (VCA) Technologies

1.1. Video Content Analysis (VCA)

1.1.1. Introduction

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.

Table 1 - Video Content Analysis 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	Used to determine the location of a camera by analyzing its output signal
Motion detection	Used to determine the presence of relevant motion in the observed scene
Object detection	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	Used in settings where the video signal has been produced. Style detection detects the style of the production process
Tamper detection	Used to determine whether the camera or its signal is tampered with.
Video tracking	Used to determine the location of persons or objects in the video signal, possibly with regard to an external reference grid

1.1.2. VCA Systems

Stand-alone and CCTV networks are present in city centers, transportation industry, counter-terror, and counter-crime infrastructures (e.g., airports, sea ports critical infrastructure, border security).

Video Content Analysis (VCA) systems are intended to improve safety and security in remote terror situations. It is an open secret that it is increasingly difficult for operators to identify a terror-related threat while viewing a bank of monitors; operators miss over 90% of such events. VCA systems assist the operators to focus on abnormal events.

Note: Video Content Analysis technology should not be confused with traditional 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:

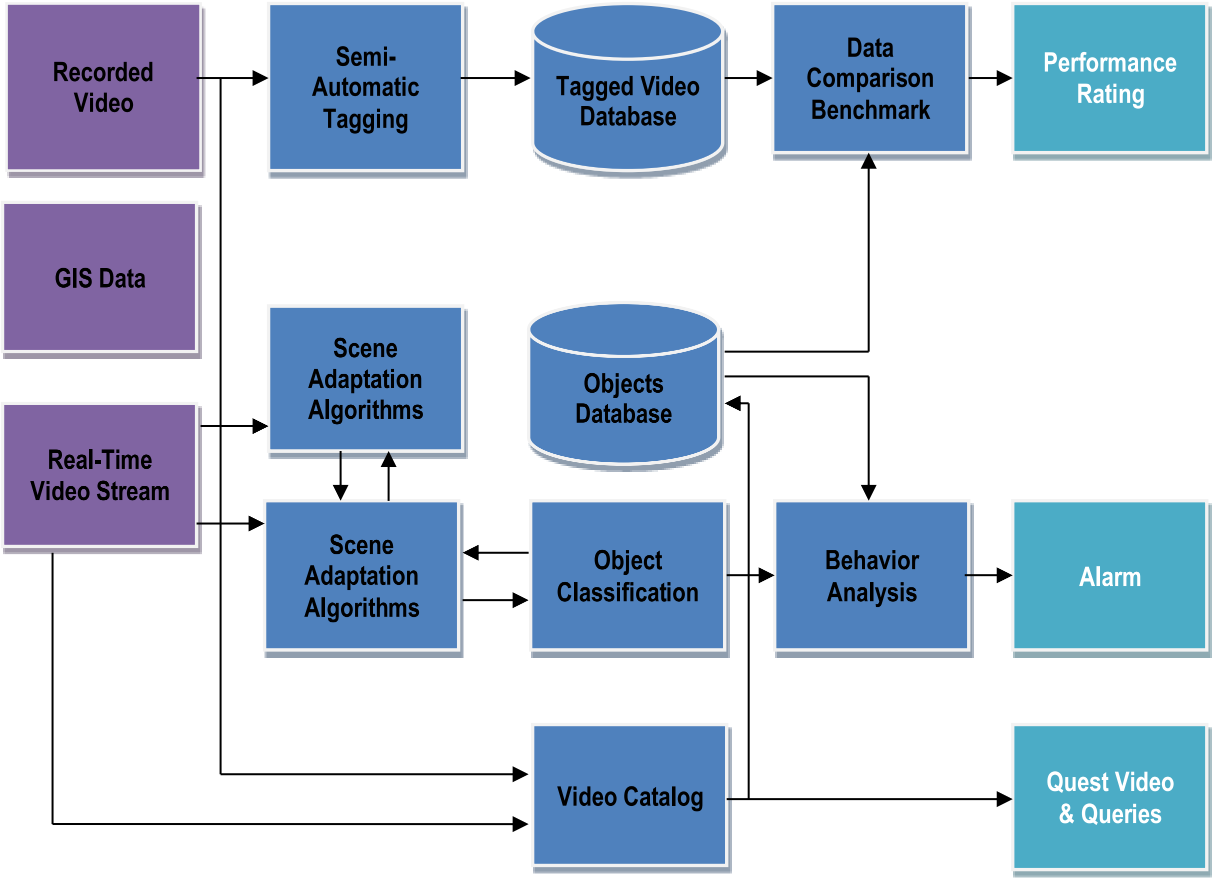
- ❑ Focus of the operator on areas of interest
- ❑ Analyzing exceptions based on statistical analysis of all the events on video.

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 issue.
- ❑ 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 exam involves blocking development of standard criteria to evaluate VCA technologies.

Figure 1 - VCA Block Diagram and Data Flow



The system typically functions in four settings:

1. Narrow outdoor field of view (e.g., perimeter security fence, runaway security)
2. Wide outdoor field of view (e.g., border checkpoint entrance area)
3. Indoor - non sterile area (e.g., airport terminal ticketing area)
4. Indoor sterile area - (e.g., airport sterile areas)

1.1.3. Pipeline VCA Research

Future VCA research and development will focus on four main issues:

Table 2 - Pipeline VCA: Sub-Systems, Technology, Technology Challenges and Target Performance

Sub-System	The Technology	Technological Challenges	Target Performance
Video Level Fusion	• Location of objects in the site • Classification • Continuous tracking	• Continuous tracking of an object around dense objects environment • Changing environmental conditions such as outdoor area • Classification of objects to varied departments around dense objects environment	• Continuous tracking within the site space (up to 10% replacement of identity) of 95% of the objects and their classification for departments (within 85% accuracy) under conditions of population density of up to 2 people per square meter, provided indoor and outdoor conditions
Learning and Behavior Analysis	• Automatic understanding of scene properties • Analysis tracks the movements of objects in space according to rules set in advance • Statistical analysis of exceptions	• Automatic parameter estimation of changing scenes • Statistical analysis of behavioral characteristics of objects at the site level, allowing exceptional analysis of the object and scene	• Adoptive software for scene conditions, without requiring manual setting of parameters • Ability to set rules defining suspects at a 95% reliability rate, and once a week, an average of false warnings of events • Creating an index (0-100) for object and scene abnormalities
Search The Catalog Video	• Maintaining and cataloguing video content according to relevant characteristics and behavior of objects • Searching databases after the fact, according to these (or similar) features – e.g.,, finding all the places where a particular object has passed, according to a specific video	• Ability to find an object in different places according to its characteristics despite the fact that every appearance of the object will look slightly different due to different camera angles and lighting conditions	• Identifying at least 75% of the different instances of each object based on a relevant query in dense environment (based on marked labels - features and/or classifications) • Tracking and monitoring capability of around 85% of the objects (i.e., the ability to identify location of any object at any given time across all the cameras
Selected Video	• Setting a basis of standard criteria for performance software and measurements ways by: • automated and semi-automated tools for building a tagged database • standard representation and transfer of information about video content	• Ability to create video tag Ground Truth in semi-automatic way so it will not be necessary to manually select each image in a video	• Creating a repository of recorded and tagged video of all relevant test scenarios around real site performance to enable measurement standard software • Standard transfer of video content (meta data)

Additional capabilities required to analyze events are:

- ❑ **Classification** – the ability to understand the type of object (person, animal, vehicle, bag, etc.)
- ❑ **Behavioral Analysis** – the ability to understand the meaning of the object/objects' movement and interaction.
- ❑ **Video Indexing** – the ability to organize video content for efficient retrieval.
- ❑ **Adaptive Software** – ability to automatically change various algorithm elements according to changes between scenes and within the scene, both to study the area and to conduct analysis.

1.1.4. Video Content Analysis Software

Intelligent CCTV surveillance systems typically use multiple video cameras, transmitting the video signals to a central control room, where a multiplex matrix is used to display certain images to security personnel. Event detection and recognition require the perceptual capabilities of human operators to detect and identify item/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 operator's 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 are increased. Automating all or part of this process would provide significant benefits, ranging from the capability to alert an operator to these 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.

1.1.5. Present Limitations of Video Analytics Software

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 software differencing software and it also allows 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, for example. 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 a technique 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 about the parameters of the movement from the telemetry system. 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 of video analytics systems is their capability in crowded scenes. A significant proportion of 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.

1.1.6. Automated Analysis Of Video Surveillance Data

In a typical system for event detection, a set of processing stages needs to be performed. First, the relevant items present on the observed scene need 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 with 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 a lot of emphasis is being put to prepare 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 CCTV surveillance processing building blocks. A general description of their purpose and known methods for accomplishing certain goals are presented.

1.1.7. Item Detection

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

One group of methods employs the so-called optical flow software which depending on clear algorithm, analyses the scene for a characteristic point 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 video flow calculation algorithm can be found in the literature making it useful for clear tasks. This approach is generally a good choice in two cases. First case is 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 compensation of 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 a lot 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, certain more sophisticated methods were proposed, among which Gaussian Mixture Models based 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 not only temporal relations in the procedure of building background models but also a number of software, involving spatio and spatio-temporal relations 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. Such an approach can be useful as in the case 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 in 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 is suitable for real time processing. Like most of the advanced methods, it introduces an adaptation procedure and therefore can be quite robust.

1.1.8. Background Subtraction: Gaussian Mixture Based Software

Detection of moving items is usually the first stage in the 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 online approximation to update the model. This technique proved to be useful in many applications as it was able to cope with illumination changes and adapt to the background model according 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. That allows modeling such events as trees swinging in the wind or traffic light sequences.

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

Item segmentation is supplemented with shadow detection and 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 wherein if 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 to see whether it belongs to a moving shadow. If the current pixel is darker than the distribution, the current pixel is assumed to be a shadow and is considered as a part of the scene background.

Another approach that is presented is where the shadow‘s shape, size, orientation, luminosity, originating position and appearance model is 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 to where the novel k-nearest neighbor pixel classifier was proposed which is applied on pixels previously classified 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 then when it starts to move again some time 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 a parked vehicle) and also causes a ghost problem. To tackle this problem, we need to compare the similarity between the edges of the detected foreground items and those of the current frame based on item level knowledge of moving items.

1.1.9. 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. A novel background estimation technique that learns the co-variation of grey levels within the incoming images using principal component analysis to generate the eigen-backgrounds has been 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 was continuously adapted to match the background conditions. As a consequence, for each incoming image, a reference frame is hypothesized efficiently to perform background segmentation from a subsample of the incoming pixels. This in turn proposes to take advantage of periodic variation of background appearance over time which is detected in the temporal frequency domain. This analysis allows estimating of 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 model is based on 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.

1.1.10. Item Detection Based on a Single-Image Algorithm

In certain scenarios, the estimation of a 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 types of software 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 items 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 like 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.

1.1.11. Item Tracking Software

Item tracking software usually follows the item detection. This building block allows for introducing 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 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 a 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 advanced limb correction building block though it also proposes graph-based particle digital filter to deal with stylistic variations of motion.

Tracking multiple items in complex backgrounds exploits the bank of Kalman filters to track each subject independently on a scene. Then, 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 about self-calibrated ground plane or depth probability density functions. This 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 creating the 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 representation, Kalman filter, scene model and a Bayesian network, to perform an on-line multi-item tracking in realistic situations from a single fixed camera.

To deal with large obstructions during tracking, this technique uses a constant acceleration motion model to track items where three predictors are employed simultaneously with a least square correlation stage to select the most likely item position. These three prediction schemas are based on α-β tracking schema.

1.1.12. Kalman Filtering Techniques, Region Segmentation

Alternatively, some people propose a technique for estimating the midpoint (or centroid) and bounding box size of each target using a Kalman digital filter with the measurements of four bounding edges. This structure facilitates the utilization of incomplete measurements that can arise due to partial occlusion.

Another challenging problem is tracking across multi camera network. The system comprises 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 with 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, in tracking it is performed in D using the Kalman digital filter together with the ground plane homographic constraint.

1.1.13. 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 is 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 that it does not rapidly change its direction and velocity of movement so 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 disappearing 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

1.1.14. Partially Observable Markov Decision Process, IVS 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 the 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. The 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 IVS 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 the system does not trigger an alarm every time a cat walks in front of the camera for example, with the speed needed to allow security staff to act on an event as quickly as possible.

For IVS 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, or 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).

Then, for any given situation if it wants to know if an intruder or object has entered the scene, the system 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 on 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, said Amato. It could also 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 said. “The POMDP model excels in decision-making regarding uncertain information, in this case whether an intruder is present or not.”

1.1.15. "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 as follows. 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 CCTV surveillance systems is not an easy task mainly due to variation of camera viewpoints. Several methods are used to solve this problem which can be divided into two general groups: feature-based and motion-based. Feature based approach utilizes the data on the item such as spatial signatures commonly related to shape or texture descriptors. One of the clear methods uses the item 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 is 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 with wavelet coefficients can be found in the literature. A particular technique based on the item 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 process 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.

1.1.16. Dimension Based Items Classifiers

Dimension based items 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's 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.

1.1.17. Shape Based Item Classifiers

Intelligent CCTV surveillance cameras can be placed variously such as 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 is a pattern recognition problem and requires a set of items for a proper classifier training.

For the purpose of classification process, Support Vector Machines were considered as a choice. 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 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).

1.1.18. Event Detection Methods

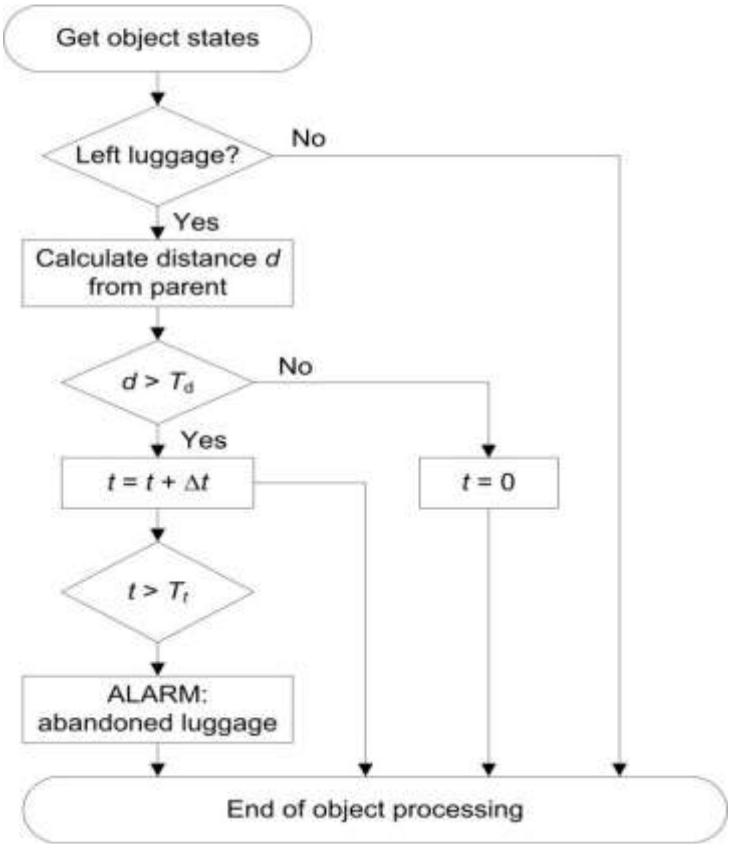
Image-processing software is used to automatically detect incidents and make measurements, relieving the control room personnel of the effort in finding out where the events take place. Events of particular interest may include abnormal stationarity, queuing, intrusion detection, loitering, unattended luggage detection as well as closely related complications such as action recognition. Certain events could be detected based on a number of predefined user rules while others require 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 and more frequent case is when a rule based approach is applied. Here, a number of atomic events like item stopped, item left the scene, 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 clearly people. In the second case, a single person or groups and crowds need to be considered. Various events can possibly be detected in each of these cases. In many cases for crowds, its flow-based clear activity and certain anomalies from the main trend can be detected. Regarding single people, primitives-based event detection as well as more complex behavior recognition can be applied.

Figure 2 - Unattended Item Detection Algorithm Flow Chart



An important problem with detection of left and removed items is that due to the nature of the background subtraction algorithm, leaving an item on 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).

1.1.19. 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 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 relating the video 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 Eigen maps 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 to perform and 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 that the subsequent inference of action label can be performed in a probabilistic manner.

1.1.20. 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 visualodometry 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 which is 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.

1.1.21. Path Reconstruction Software

Path reconstruction is one of the methods 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 are 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 specific 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:

A modified particle digital filter modeling of hidden (unobserved by cameras) places where the items are not seen in any FOV particle digital filter is chosen (instead of Kalman filter) because the process of tracking many of these items is non-linear and a certain level of noise is expected.

1.1.22. Video Cameras 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, 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 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 a monitoring system. The first is appearance of the observed item and the second is spatio-temporal data. Two-dimensional histograms are used: histogram of color for the 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 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 and others 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 a particular pair of cameras. Weights in the proposed model are estimated by "expectation maximization" algorithm. Experiments performed by the authors of this paper validate the correctness of the used method.

1.1.23. 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 about 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

The 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 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 a 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 most likely to be the one intended by the operator. The solution is recomputed 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.

More information can be found at:

Global Video Analytics, ISR & Intelligent Video Surveillance Market – 2015-2020