Video Surveillance Systems: Current status and future trends

Vassilios Tsakanikas and Tasos Dagiuklas

Division of Computer Science

Abstract

Within this report an attempted is made to document the present status of video surveillance systems. Thus, the main components of a surveillance system are presented and studied thoroughly. Algorithms for image enhancement, object detection, object tracking, object recognition and item re-identification are presented, as well as the modalities which feed these algorithms. Finally, new trends on surveillance systems, such as cloud integration, are discussed.

1. Introduction

During the past decade Video Surveillance Systems have revolved from simple video acquisition and display systems to intelligent (semi)autonomous systems, capable of performing complex procedures. Nowadays, a Video Surveillance System can integrate some of the most sophisticated, systems incorporating various types of media (e.g. sound, image, video), employing image and video analysis algorithms such as classification (e.g. neural networks or stochastic models), pattern recognition, decision making, image enhancement and several others. Thus, a modern surveillance system comprises image and video acquisition devices, data processing - analysis modules and storage units, components which are all crucial for the system's workflow[ref].

There are many variation of Video Surveillance Systems, each one trying to meet the demand of a specific market (e.g. indoor in an enterprise, public areas etc). Several categorizations can be drawn. Hence, one can categorize Video Surveillance Systems based on the type of imaging modality acquired, producing categories like "one camera systems", "many camera systems", "fixed camera systems", "moving camera systems" and "hybrid camera systems". Another categorization can be based on the applications which a Video Surveillance System can provide, such as object tracking, object recognition, ID Re-identification, customized event alerting, behavior analysis etc. Finally, Video Surveillance Systems can be categorized based on architecture a system is built on, such as stand-alone systems, cloud-aware systems and distributed systems.

For most of the time, surveillance systems have been passive and limited in scope. In this context, fixed cameras and other sensing devices such as security alarms have been used. These systems are able to track persons or to detect some kind of events (a person breaking the door or the window), however, they have not been designed to predict abnormal behaviours for instance. During the last years, there was a huge progress in sensing devices, wireless broadband technologies, high-definition cameras, and data classification and analysis. Combining such technologies in an appropriate way will allow to develop new solutions that extend the surveillance scope of the current systems and improve their efficiency.

Surveillance systems have to cope with several challenges, including, but not limited to, algorithmic, infrastructure, environmental, challenges. Thus, surveillance systems have to adapt with the emerging network and infrastructure technologies, such as cloud systems,….deep learning, video evolution (4k, HDR) in order to provide more robust and reliable services. This trend will also demand the integration of different surveillance systems for extracting more useful knowledge. This integration will require new communication protocols and data formats between surveillance agents, as well as new

surveillance adapted databases and query languages. Finally, more accurate algorithms are required, especially in the context of behavioral analysis and abnormal activities detection.

The aim of this paper is to document the current status of Video Surveillance Systems, identifing the best practices for image and video processing and analysis for surveillance systems Additionally, the applicability of proposed algorithms and architectures will be assessed, in terms of time response and scenarios variety. Moreover, open research problems and will be presented in the second part

The paper is structured as follows

3. Video Characteristics in Video Surveillance Systems

Nowadays, there is a variety of video sensors used from surveillance systems. As the technical specifications of the video sensors play a key role to the potential of a surveillance system, in this section we provide an outline of the sensors' technical characteristics.

The oldest and most used type of video sensors are analog video sensors which are used mainly for CCTV (Closed-Circuit Television) surveillance systems. The resolution of the analog cameras is measured in vertical and horizontal line dimensions and typically limited by the capabilities of both the camera and the recorder that the CCTV system is using. Table 1 highlights the common formats of analog cameras are provided, along with their resolution. Until two years ago, the higher resolution for analog systems came from the D1 format. Yet, since 2015 the AHD CCTV (Analog High Definition) cameras were introduced in the market, along with the corresponding recorders. Analog video sensors can provide frame per seconds (FPS) that canan vary between 30 FPS, down to 1 FPS. The majority of the systems use either 15 FPS or 7.5 FPS, as higher values require a large amount of storage volume, in case of recording.

Table 1: Resolutions of common analog video cameras

Analog Video format	Resolution
1080p Resolution	1920 x 1080
720p Resolution	1280 x 720
D1 Resolution	704 x 480 (NTSC for the United States) 720 x 576 (PAL for Europe)
CIF Resolution	352 x 240
QCIF Resolution	176 x 120

During the last fifteen years digital video sensors gained their market share against analog technology. While analog sensors transmit the captured data uncompressed, the digital sensors perform digitalization of the input stream and thus can take advantage of compressing algorithms and advances in video encoding. Consequentially, these sensors can interface directly with network infrastructures and transmit their data over IP-based networks. This is the reason why the digital sensors often referred as IP cameras. The resolution and the frame rate of digital sensors are adjustable. Common IP cameras, which nowadays belong to the HD (High Definition) category can capture video on 1,920 x 1,080 resolution and 30 FPS and downgrade to 1,280 x 720 or D1 for 15 FPS. Ultra HD (UHD) video sensors have been also introduced to surveillance systems, pushing the available resolutions to 4K (3840 x 2160, usually under15 FPS) or 2048x1536 under 30 FPS.

Finally, since the begging of 2010, a new type of video cameras have been introduced, the High Dynamic Range (HDR) video sensors. These sensors, which usually operates at HD resolution, are able of capturing the same scene multiple times using different exposure times (the time interval the camera shutter remains open and collects data) and then combine these frames to a single image. This

technique, which nowadays is available only to high-end video cameras, makes the bright areas of the scene darker and the dark areas brighter, enhancing the quality of the video stream. HDR cameras (as well as HD and UHD cameras) utilize the H.264 video codec.

4. Surveillance Systems and. acquired modality

All Video Surveillance systems utilize of course video streams. Yet, this is not the only modality a surveillance system can use. In this section, a brief description of systems utilizing additional modalities is provided.

4.1 Sound

The most common modality to couple with video in a surveillance system is sound. There are two types of audio-visual data fusion architectures (Cristani, Bicego, & Murino, 2007). In the first type, audio data are spatialized utilizing microphone arrays, aiming to improve tracking algorithms while in the second type, which is more general, sound is captured using a single microphone.

The most usual scenario for the first type of the systems is a known environment (indoor in the most cases), which is equipped with fixed cameras and microphones. For example, in (Checka & Wilson, 2002) and (Zotkin, Duraiswami, & Davis, 2002), moving objects are located calculating the sound time delays among the microphones. Applications using sound as modality are multi-object 3D tracking (Checka & Wilson, 2002), walking person detection (Wilson, Checka, Demirdjian, & Darrell, 2001), (Zou & B. Bhanu, Tracking humans using multi-modal fusion, 2001), (Beal, Jojic, & Attias, 2003). The approaches include audio source separation, dynamic Bayes networks, learning and interference of graphical model and 2 – layer HMM (Hidden Markov Model) frameworks.

As for the second type of fusion architectures, due to the presence of only one microphone, audio spatialization is no longer available. Hence, the most common approach for audio-visual fusion is the use of Cononical Correlation Analysis (CCA), using as variables spectral bands for sound and image pixels for video (Hardoon, Szedmak, & J.Shawe-Taylor, 2003), (Slaney & Covell, 2000). One of the main drawbacks of CCA is the need of large amount of data for model training. Some research works try to tackle this issue, like (Zou & Bhanu, Pixels that sound, 2005), in which a presumed sparsity of the audio-visual events is exploited.

Other approaches for audio-video correlation are proposed in (III J. F., Darrell, Freeman, & Viola, 2000) and (III & Darrell, 2004). According to these reports, two groups of multi-variate variables are correlated using the Maximization Mutual Information (MMI) method and in (Cuadra, Cammoun, Butz, & Thiran, 2005), where Markov chains are proposed and the audio-video joint densities are estimated using a group of training sequences. Video retrieval by content can also be facilitated by audio-video analysis. Within this context, the objects to be analyzed are typically entertainment clips (movies, commercials, news, etc.). The scope is to provide the final users the possibility to retrieve the preferred video clip from among massive amounts of visual data in a semantically meaningful and efficient manner. The heterogeneity of the sequences considered requires using high-level techniques, further heavily relying on automatic video annotation algorithms (Petkovic & Jonker, 2003), (Pfeiffer, Lienhartl, & Efflsberg, 2001).

Technology	Pros	Cons
Fixed Cameras		
Audio-Visual Fusion		

4.2 Position Systems

Video surveillance systems started to incorporate positioning data (e.g. GPS) when they incorporate moving cameras. This is accomplished by adding an extra layer of meta-data from GPS to the tracking algorithms. Yet, the raising interest for aerial video surveillance systems led to the design of surveillance architectures, which incorporated moving cameras either installed on drones or on UAV (Unmanned Aerial Vehicles). One of the first research works, which proposed a surveillance system with moving cameras was (Kumar R. , et al., 2001), where a framework for real-time, automatic exploitation of aerial video for surveillance applications is presented. The main functionality of the proposed system is performed by a module, which separates an aerial video into its natural components, namely the static background geometry, moving objects and appearance of the static and dynamic components of the scene. The system finally attempts to register the geo-location of video with the tracked objects, using GPS data and elevation maps before producing reprojected mosaics of the scenes.

Besides utilization of GPS data from UAV surveillance systems, geo-location is also used from in-vehicle surveillance systems. Systems under this framework have been proposed many research works, such as (United States Patent No. US 2008/0309762 A1, 2008) and (United States Patent No. US4789904 A, 1998). The basic idea behind these systems is the registration of the tracked objects with the GPS data, in order to facilitate the creation of a meta-data map with of the trajectories of the tracking objects.

4.3 Video

Undoubtedly, video streams are the primarily modality when it comes to surveillance systems. This is not a surprise, if we consider that vision is the sense people mostly use to explore the surrounding environment. At some level, most of the research works regarding surveillance systems try to mimic the biological process of how people detect events and categorize them. For example, a common pre-processing procedure of event detection algorithms is background/foreground classification, where the system tries to distinguish the static scene (which usually has no interest) from the dynamic foreground objects. This procedure is similar to the bioprocess where neurons detect a change in luminance and color of neighboring points after a short delay.

The quality of the acquired video stream plays a key role to the potentials of a surveillance system. Resolution, frame rate per second and contrast are some of the most important features of a video sensor. For example, a high quality video sensor can substitute a pre-processing enhancement algorithm, boosting up the response time of a surveillance system. On the other hand, usage of high resolutions results to increment of bandwidth requirements for data transmission and storage.

4.4 Modality fusion and intelligent surveillance systems

Data fusion is the process of combining two or more modalities in order to acquire more efficient and useful information compared to the acquired information when using the modalities separately. The concept of data fusion is not new, however, merging different types of data generated by heterogeneous devices is still a challenge. In the literature, different approaches to deal with this problem are suggested. Statistical analysis where typical techniques such as mean, median, standard deviation, and variance (including Kalman filtering) are used is the straightforward approach. Most of the data fusion being used now rely on probabilistic descriptions of observations and use Bayesian networks to manage the uncertainty and combine this information (Myers, Laskey, & DeJong, 1999). In this category, one can also mention the techniques based on fuzzing and Dempster-Shafer theory, and learning algorithms based on neural networks and hybrid systems. The approach to be used often depends on the type of data, the level of reliability foreseen, and the requirements of the application (in our case the intelligent surveillance) (Chen & Luo, 1999).

5. Surveillance Systems and knowledge extraction algorithms

Within this section, focus will be put on the module of a surveillance system which is responsible for "translating" the raw video data to specific structured information. The most common activities on this field are face recognition, object re-identification and object tracking.

5.1 Face recognition

Face recognition constitutes the problem of recognizing a face against a predefined knowledge database of faces. Face recognition problem implies that a face is already detected in a scene, which forces us to firstly discuss about another important problem in surveillance systems; face detection problem.

Face detection is one of the most important and well-studied problems in the computer vision literature. While the problem setup is straightforward, the algorithms proposed from the research community are numerous. The work with the largest impact in this area is (Viola & Jones, 2001). The Viola-Jones face detector comprises three main modules; namely the integral image, the classifier learning with AdaBoost and the attentional cascade structure. For the learning phase of the algorithm, several visual features have been proposed, such as Haar-like rectangular features (Mita, Kaneko, & Hori, 2005).

Face recognition has been studied from researchers for more than forty years. Various researchers have been trying to produce robust, accurate and real-time solutions. The first type of approaches have tried to model the face recognition problem as a two dimensional pattern recognition problem, calculating "important" distances of facial features, such as the distance between the eyes of the length of the lips. The second category…….

Nowadays, one can classify the methods for face recognition in three categories; namely holistic matching methods, feature-based methods and hybrid methods. According to the holistic methods, the whole face region is compared against a face database using specific techniques such as Eigenfaces (Turk & Pentland, 1991), Principal Component Analysis and Linear Discriminant Analysis (Suhas, Kurhe, & Khanale, 2012). Feature based methods are trying to extract facial geometrical features, such as mouth, lips, nose and eyes. These features are used as an input to classifiers, aiming to detect the match closest to the face detected. Feature based methods fail to produce accurate results when the aforementioned features are not visible in the scene, mainly due to unappropriated head pose. In order to tackle this problem, feature estimation methods have been proposed (Zhao, Chellappa, Phillips, & Rosenfeld, 2003), mainly taking advantage of face structural constrains. Finally, the hybrid methods take advantage of both the techniques of holistic and feature based methods. These methods use as input 3D images and for that they can use information concerning the forehead or the chin shape.

During the past few years, face recognition algorithms have come to a maturity level where they can be used on real-world applications and uncontrolled environments. This fact brought up the need for developing new approaches in face recognition problem, such as the "watch-list" problem. According to this version of the problem, the system needs to distinguish among a very large number or individuals only the people who belong in a predefined list. A research work, which tries to address this problem can be found in (Kamgar-Parsi, Lawson, & Kamgar-Parsi, 2011), where for each individual in the watch-list, a classifier is trained. Then, for the detected individuals, certain features are used as input to the classifiers, reaching to the final decision.

Face Recognition Methods	Pros	Cons
Holistic Matching		
Feature-based		
Hybrid		

5.2 ID Re-identification

The ID re-identification problem appears on multi-camera surveillance system setups, where people walk around the view field of numerous cameras. Within such setups, a surveillance system must have the ability to track people across multiple cameras, thus performing crowd movement analysis and activity detection. More specifically, given a video of a person taken from one camera, re-identification is the procedure of identifying the person from videos taken from different cameras. Re-identification is crucial in establishing reliable individuals tagging across multiple cameras or even within the same camera, when discontinuities and "blind" spots appear.

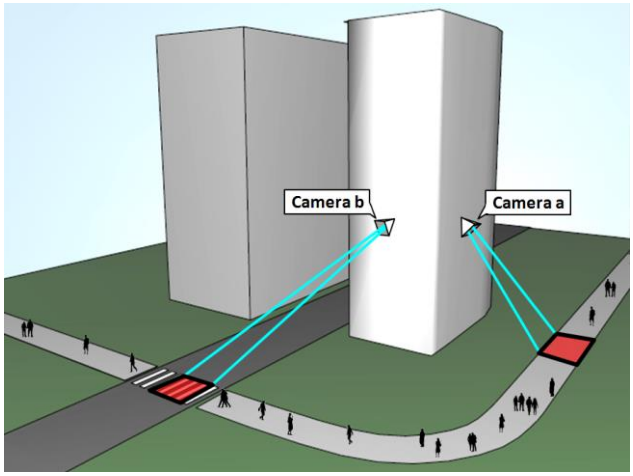

Figure 2: Person re-identification scenario[1]

ID re-identification is a challenging problem due to the visual vagueness and spatiotemporal uncertainty in a person's appearance across different cameras. These difficulties are often reinforced by either low resolution images or poor quality video streams. Issues like these forced the research community to put focus on the ID-identification problem during the last years, aiming to produce robust and wide-applicable algorithms.

Since 2010, there has been many research works, which have tried to address the ID re-identification problem. Some extensive tries can be found in (Cai & Pietikäinen, 2010), (Bazzani, Cristani, Perina, Farenzena, & Murino, 2010), (Bazzani, Cristani, & Murino, Symmetry-driven accumulation of local features for human characterization and re-identification, 2013), (Gheissari, Sebastian, & Hartley, 2006), (Wang, Doretto, Sebastian, Rittscher, & Tu, 2007). According to most of the published research studies, the problem of ID re-identification has been modeled as recognition problem. Given an image (or images) of an unknown person and a database of known people, the scope is to produce a sorted list of all the people in the database based on their similarity with the unknown individual. Thus, we expect that the highest ranked match in the database will provide an ID for the unknown person, thereby identifying the probe. In this scenario, we assume that the unknown person is included in the database of known persons (closed-set ID re-identification). Most of the approaches nowadays are based on appearance based similarity features between frames to establish common similarities. Typical features used to quantify individual similarities are low-level color features and texture features based on clothing. Nonetheless, such similarity features are only valid for a short period of time as people dress differently from day to day, or even through the same day. Hence, similarity based features are only suitable for a short period of time (short-term re-identification), which is the version of re-identification problem the research community mainly tries to solve.

[1] https://lrs.icg.tugraz.at/datasets/prid/

The methods and the techniques for ID re-identification can be categorized in several methods, as depicted in Fig. 1

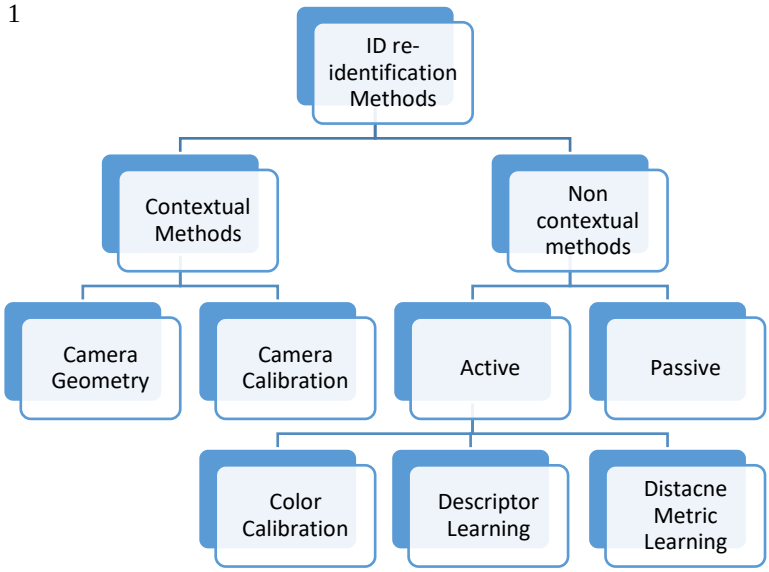

Fig. 1: ID re-identification methods categorization

Contextual methods take advantage of external information such as camera geometry and camera calibration. For example, in (Javed, Shafique, Rasheed, & Shah, 2008), camera geometry setup is taken into account in order establish intra-camera relationship and increase constrains among the cameras. Other works in this category are reported in (Makris, Ellis, & Black, 2004), (Rahimi, Dunagan, & Darrell, 2004) and (Mazzon, Tahir, & Cavallaro, 2012).

Camera topology is usually determined by correlating activities among cameras with disjoint Field of Views (Loy, Xiang, & Gong, 2010) and do not rely on information from tracking algorithms. The time delayed correlations of activities are observed and quantified, utilizing multiple camera views in a single common reference space. The estimation of the time delayed activity correlations is used for person re-identification and both spatial and temporal topology inference of a camera network.

As far as the camera calibration as context concerns, camera field of view information and homography are considered, aiming to extract features from visual descriptors. For example, (Lantagne, Parizeau, & Bergevin, 2003), individuals' height is calculated using homography, to estimate a 3D model. A Panoramic Appearance Map (PAM), which is reported in (Gandhi & Trivedi, 2007) uses information from multiple cameras that view the object to produce a single object signature. Other important works in this category are reported in (Hu, *et al.*, 2006) and in (Baltieri, Vezzani, Cucchiara, Utasi, & Szirányi, 2011). Hue *et al.* proposed a method for people tracking using multiple cameras based on the detection of principal axis for each tracking person, which are the perpendicular segments from head to toe and from shoulder to shoulder. The algorithm estimates the principle axis for each camera and then attempts to correspond them in order to re-identify people. Baltieri et al has proposed a modelling approach, where 3D information is extracted from multiple cameras. The proposed model is a 3D Markov Point Process model using two pixel-level features. The workflow includes the feature extraction from multi-plane projections of binary foreground masks and the statistical estimation of the height and the position of each person. Finally, a 3D body model based long-term tracking algorithm connects missing or hidden tracks and is used to re-identify people.

Non-contextual methods rely on knowledge extraction using only the video stream as input, ignoring the contextual data. These methods, which are reported with a high frequency during the past years are further categorized to both passive and active. Passive methods extract visual features in order to classify an individual's appearance against a known dataset (the description passive comes from the fact that these methods do not use machine learning techniques for feature extraction). Shape and color

visual features for person modelling is proposed in (Kang, Cohen, & Medioni, 2004), where the video stream is divided in polar bins and Gaussian model along with edge pixels from each bin are used to produce the features. On the same page, a spatio-temporal segmentation method, utilizing watershed segmentation is reported in (Gheissari, Sebastian, & Hartley, 2006) where the appearance of an individual is a combination of color and edge histograms. On the other hand, active methods utilize machine learning algorithms for feature extraction. A machine learning algorithm can either be supervised or unsupervised. The supervised approaches require a set of annotated training data, in order to "learn" to detect the desirable features (e.g. person's silhouette), while the unsupervised algorithms utilize clustering techniques in order to estimate different image features (without use of training data). There is another way to categorize the machine learning methods into three categories; namely distance metric learning methods, descriptor learning and calibration methods. Distance metric learning methods do not use feature selection techniques for feeding learning algorithms. Yet, they place effort on learning suitable distance metrics, which are able to maximize the accuracy of the classification, regardless of the choice of appearance representation. Typical research works of this kind are reported in (Yang & Jin, 2006) and in (Dikmen, Akbas, Huang, & Ahuja, 2010). The descriptor learning methods try to acquire the most discriminative features in order to achieve ID re-identification. Another approach is to deploy a learning phase to produce descriptive lists of features that better represent an individual's appearance using a bag-of-features approach. Such works are reported in (Wang, Doretto, Sebastian, Rittscher, & Tu, 2007), where co-occurrences between a priori learned shape and appearance features produce an individual descriptor. HOG (Histogram of Oriented Gradients) features are also utilized by many research works, such as (Dalal & Triggs, 2005) and (Zheng, Gong, & Xiang, 2009). Finally, the color calibration methods try to model the color relationships between a specific pair of cameras using color calibration techniques. They usually employ a learning phase to produce the calibration model, like the techniques reported in (Javed, Shafique, & Shah, Appearance modeling for tracking in multiple non-overlapping cameras, 2005) and in (Porikli, 2006).

When it comes to algorithms' assessment, the most common evaluation metric is the cumulative matching Characteristic Curve (CMC). This curve depicts the probability that the correct match is ranked equal to or less than a particular value against the size of the gallery set. In order to evaluate the performance of the simultaneously matching multiple probe images of the gallery, the Synthetic Re-ID Rate (SRR) curve is derived from the CMC curve, which gives the probability that any of the given fixed number of matches is correct. The normalized area under the CMC curve (nAUC) is also an important performance metric. The nAUC is the probability that the Re-ID system will produce a true match over a false (incorrect) match.

5.3 Object detection and tracking

Object detection and object tracking are the most common applications on video surveillance systems. Object detection constitutes the problem of isolating a specific region of a video stream based on the system's parameters while object tracking is a process of keeping track of the aforementioned region's motion. One can classify the object detection algorithms in four categories; namely Background Subtraction, Temporal Differencing, Frame Differencing and Optical Flow.

Algorithms using background techniques try to separate foreground objects from the background of the scene. In order to achieve this, background modeling (reference model) is mandatory. The more accurate and adaptive the background model is, the more accurate the detection algorithm is provided by (Athanesious & P.Suresh, 2012). The most common techniques to achieve background modeling include median and mead filters (M.Sankari & Meena, 2011).

Temporal Differencing algorithms calculate the difference (on pixel level) between successive video frames, in order to detect the moving objects (Joshi & Thakore, 2012). These algorithms are able to quickly adapt to highly dynamic scene changes. Yet, they suffer from important drawbacks; the most important of them is detection loss when the object stops moving and when the object's color texture is similar to the scene (camouflage) (Paragios & Deriche, 2000), (Zhu & Yuille, 1996). Also, false object detection may occur when scene objects tend to move (e.g. leaves of a tree when the air is blowing).

A simple approach of temporal differencing is Frame Differencing, where the temporal information indicates the moving objects of the scene. In such methods, presence of mobility is established by calculating the difference (pixel level) of two successive video frames (S.Rakibe & D.Patil, 2013).

Finally, Optical flow is the pattern of objects motion in a visual scene caused by the relative motion between an observer and the scene. Optical flow methods use partial derivatives with respect to the spatial and temporal coordinates in order to calculate the motion between two image frames. Such methods seem to be more accurate that the aforementioned approaches, but due to the computational time required and the noise tolerance, they are unsuitable for real (or near real) time scenarios.

Regarding the object tracking algorithms, their scope is to return the route for an object by calculating its relative position for each video frame (Zhang & Ding, 2012). Object tracking can be classified as point based tracking, kernel based tracking and silhouette based tracking (Athanesious & P.Suresh, 2012) (see Fig. 2).

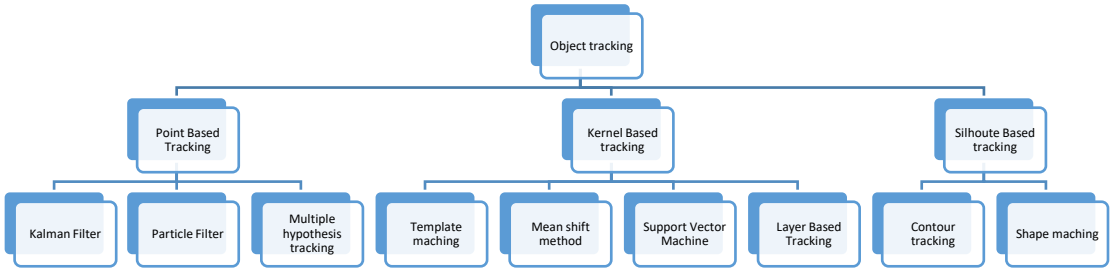

Fig 2: Object tracking methods

The most common point-based approaches utilize Kalman and Particle filters. Kalman filter (Welch & Bishop, 2006) is a set of equations that provide recursive computational means to estimate a process's past, presence and future. Methods utilizing Kalman filter are based on Optimal Recursive Data Processing Algorithm. On the other hand, Particle Filter (Athanesious & P.Suresh, 2012) generates all models for one variable (e.g. contours, color features, or texture mapping). The particle filter is actually a Bayesian sequential importance technique. In Multiple Hypothesis Tracking algorithm, several frames are observed for better tracking outcomes (iteration algorithm). Each hypothesis is a crew of disconnect tracks and for each hypothesis, an estimation of object's position in the following frame is made. The predictions are then compared by calculating a distance measure, allowing multiple hypothesis tracking algorithm to track multiple objects.

In Kernel based tracking, kernel refers to the object representations of rectangular or ellipsoidal shape and object appearance. The objects are tracked by estimating the movement of the kernel on each successive frame (Athanesious & Suresh, 2013). Kernel based approaches can be classified in four categories, namely template matching, mean shift method, support vector machine and layer based tracking. Template matching algorithms (Athanesious & Suresh, 2013), (Saravanakumar, Vadivel, & Ahmed, 2010) employ a brute force method for the examination of the video frame, aiming to detect the region of interest. In template matching, a reference image is verified with the frame that is separated from the video. Template matching algorithms are able to detect small pieces of a reference image, but the usually work for only one object and they require computational heavy load. The second category of the kernel based methods is the Mean Shift Method. The Mean Shift algorithm aims to detect the region of a frame that is most similar to a reference model. For modeling, either the reference object or the "key" object, probability density functions are utilized as well as color histograms. Support Vector Machines (SVM), the third category of kernel based approaches, is a wide used classification scheme. According to these algorithms, each sample (usually pixel groups) of a video

frame is classified as either “tracking object” or “non-tracking object” (Mishra, Chouhan, & Nitnawwre, 2012). Such approaches can handle partial occlusion of the tracking object but they require a training phase. Finally, according to the Layering based tracking, each frame is separated to three layers; namely, shape representation (ellipse), motion (such as translation and rotation,) and layer appearance (based on intensity). Such approaches can handle tracking of multiple objects.

Concluding with the object tracking algorithms, the Silhouette Based Tracking approaches are discussed. These algorithms are used to track objects with complex shapes, such as fingers. Silhouette based methods utilize accurate shape descriptions for the objects. Silhouette based tracking approaches are categorized as either contour tracking methods (Athanesious & Suresh, 2013), where a contour reshapes from frame to frame aiming to keep track with the object or Shape Matching algorithms, where only one frame is examined from time to time (without knowledge passed from the previous frame), using density functions, silhouette boundary and object edges (Athanesious & P.Suresh, 2012).

6. Quality Enhancement Algorithms

The knowledge extraction algorithms discussed in the previous section use as input frames or video streams, in order to either enhance the quality of the modalities or to provide an initial layer of information for the next processing level. In this section, we will discuss some of the most important quality enhancement methods as well as the most common preprocessing algorithms.

6.1 Foreground/background identification

Foreground/background modeling identification is the process where each pixel of a scene is classified in two classes; either F (denoting the foreground) or B (denoting Background), which can be eliminated to a one-class classification problem. Foreground includes the surveillance subject while the background includes the rest of the scene. There are several approaches, which can model the background, as depicted in Fig. 3.

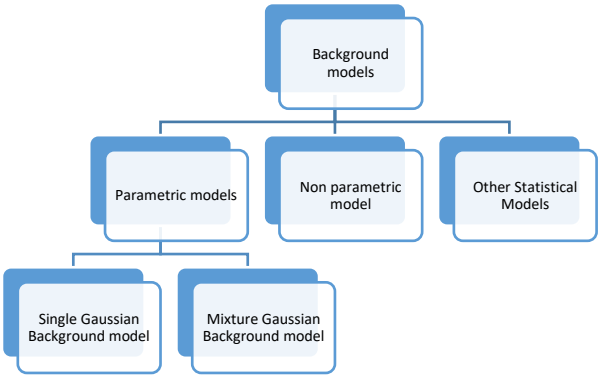

Fig. 3: Background modelling approaches

According to Single Gaussian background models, the noise distribution at a given pixel can be modeled as a zero mean Gaussian distribution. Thus, the intensity (or any other pixel feature) at a pixel is a random variable with a Gaussian distribution (Forsyth & Ponce, 2002), (Gao & Boult, 2000). In the case of colorful images, a multivariate Gaussian model is used. This model can be adaptive to slow changes in the scene (e.g. dust) by recursively updating the mean with each new frame. Single Gaussian Background models fail to model (usually) outdoor environments, where background is not static (e.g. leaves of a tree). In order to model such scenes, a generalization based on a Mixture of Gaussians has been proposed. This model was introduced by (Friedman & Russell, 1997), while (Stauffer & Grimson, 1999) have proposed a generalization to the previous approach.

The need of modeling highly dynamic scenes requires a much more flexible background modelling. This leaded to the use of non-parametric density estimator for background modeling (Elgammal, Harwood, & Davis, 2000). All non-parametric density estimation methods (e.g. histograms, ..) are asymptotically kernel methods (Scott, 1992), and a wide used non-parametric model is the kernel density estimation technique, which estimates the underlying density and is quite general (Duda, Stork, & Hart, 2000).

Lastly, in the literature there have been proposed other statistical techniques for background modelling. For example, in (Toyama, Krumm, Brumitt, & Meyers, 1999) linear prediction using the Wiener filter is proposed to predict pixel intensity given a recent history of pixel values. Linear prediction using the Kalman Filter was used in (Karmann, Brandt, & Gerl, 1990) and in (Koller, et al., 1994). Another approach has applied Hidden Markov Models to model a wide range of variations in the pixel intensity. These variations are modeled as discrete states corresponding to modes of the environment, for example cloudy vs. sunny (Patwardhan, Sapiro, & Morellas, 2008). Other approaches utilize background subtraction techniques, which deal with quasi-moving background, e.g. scenes with dynamic textures. One robust algorithm of this approach is an Auto Regressive Moving Average model (ARMA) (Soatto, Doretto, & Wu, 2001), where a Kalman filter was used in order to update the model. Finally, a biologically-inspired non-parametric background subtraction approach was proposed in (Maddalena & Petrosino, 2008), where the pixel process is modeled as an artificial neural network.

As far as the features that are used for Background Modeling concern, intensity has been the most commonly-used feature. Alternatively, many research works report edge features. The use of edge features for background modelling is inspired by the need to have a brightness invariant representation of the scene. Another feature used is optical flow (A. Mittal & Paragios, 2004), which was used to capture background dynamics. Apart from pixel-based approaches, block-based approaches have also been used for background modeling. For example, block matching has been used for detection of changes between successive frames (Hsu, Nagel, & Rekers, 1984).

6.2 Image/Video quality enhancement algorithms

Image/Video enhancement algorithms are mandatory for any surveillance system. This is due to the fact that low quality sensors and multivariate environmental conditions (e.g. fog, rain) produce highly noisy video streams. Hence, enhancement algorithms are crucial for the robust function of applications such as object detection and object tracking. There are two main techniques for image processing depending on the domain each technique works; namely spatial based and frequency-based domain. Spatial based domain refers to the image plane itself, and algorithms in this class process the image pixels directly while frequency-based domain processing techniques represent the image in the spatial domain and manipulate the spatial frequency spectrum. Research community has proposed several methodologies for improving the quality of image/video input, which can be categorized as shown in Fig. 4.

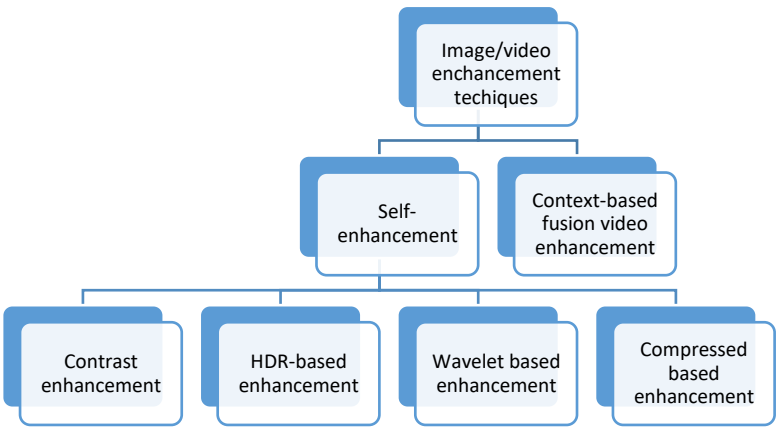

Fig. 4: Categories of image/video enhancement techniques

Self enhancement techniques refer to the techniques that use as input only the image/video under examination. There are four categories in this class. The first category refers to modifications on the contrast map of an image. The aim is to adjust the local contrast in different regions of the image so that the "hidden" details in shady or bright regions are revealed. There are numerous algorithms for contrast enhancement which all aim at taking advantage the parts of the dynamic range that are "inactive". Widely used algorithms are power low rule, gamma manipulation, histogram equalization and tone mapping. Power Low Rule, gamma manipulation. Histogram equalization aim to uniformly distribute an image's histogram utilizing density functions (Du & Ward, 2010). On the other hand, tone mapping techniques take under consideration the display device of video, trying to map the tone between the video input and the tone of the display device (Reinhard, Ward, Pattanaik, & Debevec, 2005). HDR-based enhancement techniques are the second category of self-enhancement methods. High dynamic range imaging (HDR) is a set of methodologies that offer a larger dynamic range of brightness between the brightest and the darkest pixel. HDR images can be produced by either combing multiple images of the same scene taking under different exposure values (Heo, Lee, Lee, SMoon, & Cha, 2011) or by using image processing algorithms such as the one proposed in (Petro, Sbert, & Morel, 2014). The third category utilizes wavelet transformation, producing a wavelet image, suitable for processing the image/video. The wavelet techniques utilize wavelet coefficients, wavelet shrinkage denosing or the dual-tree complex wavelet transform (Wan, Canagarajah, & Achim, 2007). Finally, the compressed based enhancement algorithms operate directly on the transform coefficients (e.g. Discrete Cosine Transform) of the images that are compressed. As far as the context – based fusion enhancement techniques concern, they utilize information from other modalities, or even from past data of the same sensing device in order to overcome poor light conditions and other environmental noisy situations (Asmare, Asirvadam, Iznita, & Hani, 2010).

6.3 Limitations

All of the aforementioned algorithms and techniques are innovative and provide solutions to by any means non trivial problems. Yet, almost all of the approaches share, more or less, the same weaknesses. First of all, while the majority of video processing algorithms (such as motion detection) work fairly well, when we move to video analysis algorithms (such as human running detection), the response time of the systems increase and the accuracy decreases. Additionally, as debated in (Porikli, et al., 2013), most of the test databases used to evaluate the performance of surveillance systems don't include heterogeneous datasets. Thus, the documented accuracy of proposed algorithms differs, sometimes to a great extent, when they are tested to real life scenarios, where the lighting and weather conditions are constantly changing.

Thus, there is a great need of designing approaches which will be more robust and more reliable, increasing the applicability and therefor the economy scale of surveillance systems.

7. Surveillance Systems and computing infrastructures

7.1 Cloud architectures

Real time or near-real time response is perhaps the most important factor when it comes to surveillance systems. Automatic alerting upon a specific event is only valuable when it occurs within a time window after the actual event. Nowadays, surveillance systems, which meet the aforementioned requirement have been designed and deployed all around the world. Yet, the nature of the events which are recognized automatically from the systems are rather trivial, including object moving, fire existence or object recognition.

Nonetheless, surveillance systems face today a set of challenges, which involve car accidents detection, terrorist activities prediction or multipurpose behavioral analysis. These events require a substantial

larger computational resources, as they comprise complex calculations and non-linear models. On top of this, modern video sensors are able to capture HD and HDR footages, which facilitate the event detection algorithms and tackles, to a certain point, bad lighting conditions and other artifacts (Boschetti, Adami, Leonardi, & Okuda, 2011). The result of incorporating such sensors into surveillance systems is the proliferation of the produced data rates and of course the increment of the required storage size.

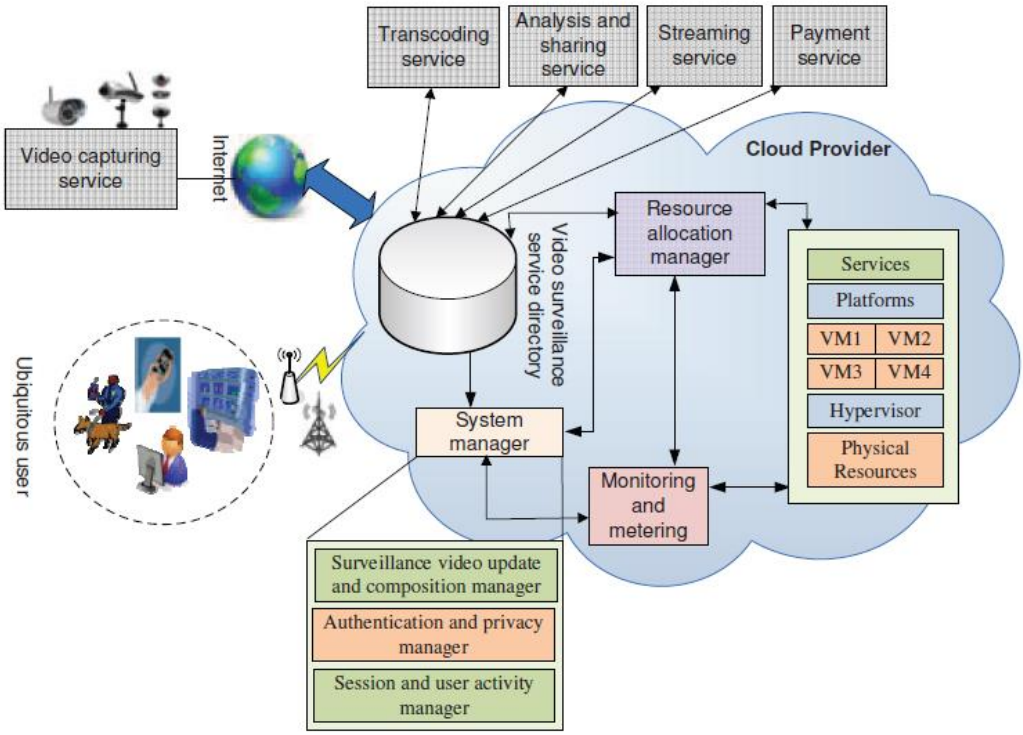

Fig: Integration of Surveillance with Cloud-based infrastructures

Both requirements for additional computational capabilities and storage size increment could be addressed by integrating surveillance systems with cloud infrastructures. As promising as this possibility sounds, there are not many reported surveillance systems in the literature, which use cloud services, either as SaaS (Software as a Service), as PaaS (Platform as a Service) or as IaaS (Infrastructure as a service). One of these works is documented in (Wu, Chang, Juang, & Yen, 2012), where the architecture and the operation flow of a video surveillance system which utilize cloud and P2P technologies, namely Hadoop and MapReduce. Based on their design, this approach can provide scalability, efficiency and reliability to a surveillance system. A similar work is reported in (Rodríguez-Silva, Adkinson-Orellana, González-Castaño, Armiño-Franco, & González-Martínez, 2012), where the proposed cloud infrastructure is used as SaaS and focus mainly on storage issues, using Amazon S3 platform. On the same track, (Li, Zhang, & Yu, 2011) describe a surveillance system for urban traffic systems, which is able to process massive floating car data coming from city taxis. Bigtable and MapReduce are explored as cloud technologies for not only storage purposes but also for computational processes. Finally, a resource allocation scheme for service management in cloud-based surveillance systems is described in (Hossain, Mehedi Hassan, Al Qurishi, & Alghamdi, 2012), where VM (Virtual Machines) resources are tuned based on QoS requirements, as depicted in Figure 3.

Figure 3: Proposed conceptual cloud architecture (Hossain, Mehedi Hassan, Al Qurishi, & Alghamdi, 2012)

8. Surveillance Systems and Augmented Reality

Augmented Reality, in the context of surveillance systems, refer to the information depicted on the operator's screen(s) on top of the video stream captured by the surveillance cameras of the system (Ismail , Hu, You, & Neumann, 2003). The type of the projected information range from static information to object tracking trajectories, dynamic labeling of detected objects and missing or hidden objects.

Some of the most important studies on this field come from surveillance system used for military purposes. For instance, in (Hall & Trivedi, 2002) a scheme is proposed for observing multiple video streams, while in (Spann & Kaufman, 2000), a visualization system is proposed by merging dynamic imagery with geometry model of a battlefield visualization. In (Kumar, Sawhney, Guo, Hsu, & Samarasekera, 2000), an augmented visualization of urban locations is reconstructed using offline video streams and 3D location models. Finally, a system which automatically detects humans and vehicles from multiple video streams and then extract and place selected frames on a map, thus reducing the workload of the operator, is described in (Kanade, Collins, Lipton, Burt, & Wixson, 1998).

9. Future trends on Video Surveillance systems

During the past three decades, an enormous set of works addressing the problem of automatic (or semi-automatic) surveillance has been proposed by the research community. Main subtasks that were studied were object tracking, object re-identification, object recognition and image enhancement. Within this framework, many excellent studies have proposed algorithms and systems which address the aforementioned problems with (more than) acceptable accuracy and robustness.

Figure: Research trends of surveillance systems

Yet, there are several research challenges. Most of the video surveillance systems seem to share two common limitations. The first limitation refers to a (too) high false alarm rate in detection of interesting events within the surveillance scene. This drawback causes various problems to the owners of the surveillance systems and they usually decide to deactivate automatic alerting features. Secondly, existing surveillance systems fail to function properly under all real-world conditions, such as rain, fog, snow, blowing dust, water on the lens or image plane artifacts.

In order to overcome the aforementioned limitations, new algorithms and techniques need to be developed, increasing the accuracy and the robustness of the surveillance systems. Besides addressing flaws of already established surveillance systems, researchers working on video analytics should bring surveillance to the next level, working on the following topics.

A. **Cloud infrastructures integration**. Cloud technology seems to match perfectly with surveillance systems, as it can offer both the missing computational power video analytics require and the storage capacity usually a surveillance system needs. Cloud infrastructures are expected to facilitate installation and management of surveillance systems, shifting the paradigm from standalone applications to Software-as-a-Service. This will allow surveillance systems to use different video analytics and alerting mechanisms when it is required and for the time period it is required. Bearing in mind the cost transmitting a video footage to a cloud system and the cost of cloud storage, new compression algorithms need to be designed, which will maintain the accuracy of the video analytics algorithms while reducing the aforementioned costs.
B. **Edge Computing and Fog Computing**
C. **Communication protocols between surveillance systems.** Despite the fact that surveillance systems become more and more popular, there is no specific protocol for communication between them. Such protocols would be extremely useful for public safety scenarios and terrorism prevention, facilitating information exchange between different surveillance systems deployed around a city. Thus, analytics such as object re-identification and object tracking would be possible among different and heterogeneous surveillance systems.
D. **Multi-Modality fusion**. Apart from video, which is the dominating sensing technology for surveillance systems, other modalities can facilitate monitoring and alerting tasks. Such

modalities include audio, thermal video, night vision video, HDR video and GPS tags. Thus, algorithms and techniques are required, in order not only to seamlessly fuse these modalities to a single output but also to automatically decide which modalities are more suitable for either different conditions or for different tasks. These approaches, among other applications, are expected to provide to autonomous vehicles (such as drones) the functionality of "deciding" which sensors are more appropriate to use on different situations.

E. **Analytics and scene reasoning**. The ultimate aim of an intelligent surveillance system is to automatically produce high-level information of the recorded scene, such as objects identification and motion recognition. Other tasks, such as tracking of individual people in crowds, keeping track of moving objects that are temporally occluded, and tracking and understanding interactions between multiple targets are further challenges that aren't yet reliably addressed. While the research community has proposed an extended set of algorithms and techniques in this area, higher levels of accuracy and applicability are required.

F. **Surveillance databases & event oriented query languages.** The usual scenario of a surveillance system is to store the video footage for a pre-defined time-frame in order to use it in case of a future events, related to the area under surveillance. In such scenarios, the common practice is to review the video streams which is a rather time-consuming and resource demanding task. As we use surveillance systems to capture events, surveillance databases must be event oriented, improving not only the workflow of a person seeking a specific event, but also the storage capacity of a system, as we will avoid the pointless saving of the whole video footage and focus on storing the events. Such databases will be integrated with event oriented query languages, in order to facilitate seeking tasks and high level knowledge extraction tasks.

G. **Augmented reality on surveillance systems.** Offering in real time (or in near-real time) information, analytics and metadata about a monitoring scene would undoubtedly help surveillance operators to work with several monitors and with crowded scenes. Thus, producing virtual reality information and over layering it with the actual video footage is a challenging task that needs to be further addressed. Additionally, generating an auditory display for complex scenes is very appealing to support situational awareness in surveillance systems. Approaches like these are expected to improve the workflow of monitoring

For addressing the research directions mentioned above, several issues must be resolved in parallel. One of the most important is the real time response of a surveillance system. For this, innovative cloud infrastructures are expected to provide the surveillance systems with the appropriate computational power and storage space. Finally, new video sensors, such as UHD and HDR cameras are expected to feed the surveillance systems with quality video streams, reducing the necessity of image enhancement and preprocessing algorithms.

10. Conclusion

Within the report we attempted to provide an overview of the video surveillance systems today, both from the algorithmic and from the systemic point of view. Video surveillance systems are expected to play a key role in future smart houses and future cities providing an innovative set of services to end users. Automation and real-time response are expected to allow the deployment of video surveillance systems to millions of square meters, multiplying the economics of the area.

11. Bibliography

A. Mittal, A., & Paragios, N. (2004). Motion-based background subtraction using adaptive kernel density estimation. *CVPR*, 302–309.

al., H. e. (2008). *United States Patent No. US 2008/0309762 A1*.

Asmare, M. H., Asirvadam, V. S., Iznita, L., & Hani, A. F. (2010). Image enhancement by fusion in contourlet transform. *International Journal on Electrical Engineering and Informatics, 2*(1), 29-42.

Athanesious, J., & P.Suresh. (2012). Systematic Survey on Object Tracking Methods in Video. *International Journal of Advanced Research in Computer Engineering & Technology*, 242-247.

Athanesious, J., & Suresh, P. (2013). Implementation and Comparison of Kernel and Silhouette Based Object Tracking. *International Journal of Advanced Research in Computer Engineering & Technology*, 1298-1303.

Baltieri,]. D., Vezzani, R., Cucchiara, R., Utasi, C. B., & Szirányi, T. (2011). Multi-view people surveillance using 3D information. *The Eleventh International Workshop on Visual Surveillance (in conjunction with ICCV 2011)*, 1817–1824.

Bazzani, L., Cristani, M., & Murino, V. (2013). Symmetry-driven accumulation of local features for human characterization and re-identification. *Comput. Vis. Image Underst., 117*(2), 130–144.

Bazzani, L., Cristani, M., Perina, A., Farenzena, M., & Murino, V. (2010). Multiple-shot person re-identification by hpe signature. *International Conference on Pattern Recognition*, (pp. 1413–1416).

Beal, M., Jojic, N., & Attias, H. (2003). A graphical model for audiovisual object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 828–836.

Boschetti, A., Adami, N., Leonardi, R., & Okuda, M. (2011). An optimal video-surveillance approach for HDR videos tone mapping. *EUSIPCO*, (pp. 274-277). Barcelona, Spain.

Cai, Y., & Pietikäinen, M. (2010). Person re-identification based on global color context. *The Tenth International Workshop on Visual Surveillance (in conjunction with ACCV 2010)*, (pp. 205–215).

Checka, N., & Wilson, K. (2002). Person Tracking Using Audio-Video SensorFusion. *MIT Artificial Intelligence Laboratory*.

Chen, T. M., & Luo, R. C. (1999). Multilevel Multiagent Based Team Decision Fusion for Autonomous Tracking System. *Machine Intelligence & Robotic Control, 1*(2), 63-69.

Cristani, M., Bicego, M., & Murino, V. (2007). Audio-Visual Event Recognition in Surveillance Video Sequences. *IEEE Transactions on Multimedia, 9*(2), 257 - 267.

Cuadra, M. B., Cammoun, L., Butz, T., & Thiran, J. C. (2005). From error probability to information theoretic (multi-modal) signal processing. *Signal Processing, 85*(5), 875–902.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 886–893.

Dikmen, M., Akbas, E., Huang, T., & Ahuja, N. (2010). Pedestrian recognition with a learned metric. *Asian Conference in Computer Vision*, 501–512.

Du, S., & Ward, R. K. (2010). Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions. *IEEE Trans. Circuits and Systems for Video Technology, 99*, 1-12.

Duda, R., Stork, D., & Hart, P. (2000). *Pattern Classification.* New York: Wiley.

Elgammal, A., Harwood, D., & Davis, L. (2000). Nonparametric background model for background subtraction. *6th European Conference of Computer Vision.*

Forsyth, D., & Ponce, J. (2002). *Computer Vision a Modern Approach.* Upper Saddle River: Prentice Hall.

Friedman, N., & Russell, S. (1997). Image segmentation in video sequences: a probabilistic approach. *Uncertainty in Artificial Intelligence*.

Gandhi, T., & Trivedi, M. (2007). Person tracking and reidentification: introducing panoramic appearance map (pam) for feature representation. *Mach. Vis. Appl., 18*, 207–220.

Gao, X., & Boult, T. (2000). Error analysis of background adaption. *IEEE Conference on Computer Vision and Pattern Recognition*.

Gheissari, N., Sebastian, T., & Hartley, R. (2006). Person reidentification using spatiotemporal appearance. *IEEE Conference on Computer Vision and Pattern Recognition*, 1528–1535.

Grimson, W., Stauffer, C., & Romano, R. (1998). Using adaptive tracking to classify and monitor activities in a site. *IEEE Conference on Computer Vision and Pattern Recognition*.

Hall, B., & Trivedi, M. (2002). A novel graphical interface and context aware map for incident detection and monitoring. *9th World Congress on Intelligent Transport Systems.*

Hardoon, D., Szedmak, S., & J.Shawe-Taylor. (2003). Canonical correlation analysis an overview with application to learning methods. *Tech. Rep. CSD-TR-03-02.*

Heo, Y. S., Lee, K. M., Lee, S. U., SMoon, Y., & Cha, J. Y. (2011). Ghost-free high dynamic range imaging. *Proc. of Asian Conference on Computer Visio*, 486-500.

Hossain, M., Mehedi Hassan, M., Al Qurishi, M., & Alghamdi, A. (2012). Resource Allocation for Service Composition in Cloud-based Video Surveillance Platform. *2012 IEEE International Conference on Multimedia and Expo Workshops.*

Hsu, Y. Z., Nagel, H., & Rekers, G. (1984). New likelihood test methods for change detection in image sequences. *Comput. Vis. Image Process., 26*, 73–106.

Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., & Maybank, S. (2006). Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. Pattern. Anal. Mach. Intell., 28*, 663–671.

III, J. F., & Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 406–413.

III, J. F., Darrell, T., Freeman, W., & Viola, P. (2000). Learning joint statistical models for audio-visual fusion and segregation. *Advances in Neural Information Processing Systems*, 772–778.

Ismail , S. O., Hu, J., You, S., & Neumann, U. (2003). 3D Video Surveillance with Augmented Virtual Environments. *IWVS '03 First ACM SIGMM international workshop on Video surveillance*, (pp. 107-112). Berkeley, California.

Javed, O., Shafique, K., & Shah, M. (2005). Appearance modeling for tracking in multiple non-overlapping cameras. *IEEE Conference on Computer Vision and Pattern Recognition, 2*, 26–33.

Javed, O., Shafique, K., Rasheed, Z., & Shah, M. (2008). Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Comput. Vis. Image Underst, 109*(2), 146–162.

Joshi, K. A., & Thakore, D. G. (2012). A Survey on Moving Object Detection and Tracking in Video Surveillance System. *International Journal of Soft Computing and Engineering, 2*(3), 2231-2307.

Kamgar-Parsi, B., Lawson, W., & Kamgar-Parsi, B. (2011, Octomber). Toward Development of a Face Recognition System for Watchlist Surveillance. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 33*(10), 1925-1937.

Kanade, T., Collins, R., Lipton, A., Burt, P., & Wixson, L. (1998). Advances in cooperative multi-sensor video surveillance. *Proc. of DARPA Image Understanding Workshop.* DARPA.

Kang, J., Cohen, I., & Medioni, G. (2004). Object reacquisition using invariant appearance model. *Proceedings of International Conference on Pattern Recognition, 4*, 759–762.

Karmann, K. P., Brandt, A. V., & Gerl, R. (1990). Moving object segmentation based on adabtive reference images. *Signal Processing V: Theories and Application*.

Koller, D., Weber, J., Huang, T., Malik, J., Ogasawara, G., Rao, B., & Russell, S. (1994). Towards robust automatic traffic scene analyis in realtime. *International Conference of Pattern Recognition.*

Kumar, R., Sawhney, H., Guo, Y., Hsu, S., & Samarasekera. (2000). 3D manipulation of motion imagery. *ICIP*.

Kumar, R., Sawhney, H., Samarasekera, S., Hsu, S., Tao, H., Guo, Y., . . . Burt, P. (2001). Aerial Video Surveillance and Exploitation. *PROCEEDINGS OF THE IEEE, 89*(10), 1518-1539.

Lantagne, M., Parizeau, M., & Bergevin, R. (2003). Vip: vision tool for comparing images of people. *Vision Interface*, 35–42.

Li, Q., Zhang, T., & Yu, Y. (2011). Using cloud computing to process intensive floating car data for urban traffic surveillance. *International Journal of Geographical Information Science, 25*(8), 1303–1322.

Loy, C., Xiang, T., & Gong, S. (2010). Time-delayed correlation analysis for multi-camera activity understanding. *Int. J. Comput. Vis., 90*(1), 106–129.

M.Sankari, & Meena, C. (2011). Etimation of Dynamic Background and Object Detection in Noisy Visual Surveillance. *International Journal of Advanced Computer Science and Applications*, 77-83.

Maddalena, L., & Petrosino, A. (2008). A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Trans.Image Process., 17*(7), 1168–1177.

Makris, D., Ellis, T., & Black, J. (2004). Bridging the gaps between cameras. *IEEE Conference on Computer Vision and Pattern Recognition, 2*, 205–210.

Mazzon, R., Tahir, S., & Cavallaro, A. (2012). Person re-identification in crowd. *Pattern Recogn. Lett., 33*(14), 1828–1837.

Mishra, R., Chouhan, M. K., & Nitnawwre, D. (2012). Multiple Object Tracking by Kernel Based Centroid Method for Improve Localization. *International Journal of Advanced Research in Computer Science and Software Engineering*, 137-140.

Mita, T., Kaneko, T., & Hori, O. (2005). Joint Haar-like features for face detection. *In Proc. of ICCV*.

Myers, J. W., Laskey, K. B., & DeJong, K. A. (1999). Learning Bayesian Networks from Incomplete Data using Evolutionary Algorithms. *Proceedings of The Genetic and Evolutionary Computation Conference.* Orlando.

Paragios, N., & Deriche, R. (2000). Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. Pattern Analysis Machine Intelligence, 22*(3), 266–280.

Patwardhan, K., Sapiro, G., & Morellas, V. (2008). Robust foreground detection in video using pixel layers. *IEEE Trans. Pattern Anal. Mach. Intell., 30*, 746–751.

Peterson, R. D. (1998). *United States Patent No. US4789904 A.*

Petkovic, M., & Jonker, W. (2003). Content-Based Video Retrieval : A Database Perspective. *Multimedia Systems and Applications*.

Petro, A. B., Sbert, C., & Morel, J.-M. (2014). Multiscale Retinex. *Image Processing On Line*.

Pfeiffer, S., Lienhartl, R., & Efflsberg, W. (2001). Scene determination based on video and audio features. *Multimedia Tools Appl, 15*(1), 59–81.

Porikli, F. (2006). Inter-camera color calibration by correlation model function. *International Conference on Image Processing, 2*, II–133–6.

Porikli, F., Brémond, F., Dockstader, S. L., Ferryman, J., Hoogs, A., Lovell, B. C., . . . Venetianer, P. L. (2013). Video Surveillance: Past, Present, and Now the Future. *IEEE SIGNAL PROCESSING MAGAZINE*, 190-199.

Rahimi, A., Dunagan, B., & Darrell, T. (2004). Simultaneous calibration and tracking with a network of non-overlapping sensors. *IEEE Conference on Computer Vision and Pattern Recognition, 1*, 187–194.

Reinhard, E., Ward, G., Pattanaik, S., & Debevec, P. (2005). *High dynamic range imaging: acquisition, display, and image-based lighting.* San Francisco, CA: Morgan Kaufmann Publishers Incorporated.

Rodríguez-Silva, D., Adkinson-Orellana, L., González-Castaño, F., Armiño-Franco, I., & González-Martínez, D. (2012). Video surveillance based on cloud storage. *2012 IEEE Fifth International Conference on Cloud Computing.*

S.Rakibe, R., & D.Patil, B. (2013). Background Subtraction Algorithm Based Human Motion Detection. *International Journal of Scientific and Research Publications*.

Saravanakumar, S., Vadivel, A., & Ahmed, C. S. (2010). Multiple human object tracking using background subtraction and shadow removal techniques. *2010 International Conference on Signal and Image Processing*, (pp. 79-84).

Scott, D. (1992). *Mulivariate Density Estimation.* Hoboken: Wiley-Interscience.

Slaney, M., & Covell, M. (2000). Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. *Proc. Neural Information Processing*.

Soatto, S., Doretto, G., & Wu, Y. (2001). Dynamic textures. *International Conference on Computer Vision, 2*, pp. 439–446.

Spann, J., & Kaufman, K. (2000). Photogrammetry using 3D graphics and projective textures. *IAPRS.*

Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. *IEEE Conference on Computer Vision and Pattern Recognition*.

Suhas, S., Kurhe, A., & Khanale, P. (2012). Face Recognition Using Principal Component Analysis and Linear Discriminant Analysis on Holistic Approach in Facial Images Database. *IOSR Journal of Engineering, 2*(12), 15-23.

Toyama, K., Krumm, J., Brumitt, B., & Meyers, B. (1999). Wallflower: principles and practice of background maintenance. *IEEE International Conference on Computer Vision.*

Turk, M. A., & Pentland, A. P. (1991). *Face Recognition Using Eigenfaces.*

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In Proc. of CVPR*.

Wan, T., Canagarajah, N., & Achim, A. (2007). Multiscale color-texture image segmentation with adaptive region merging. *Proc. of IEEE International Conference on Acoustics*, I-1213 - I-1216.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., & Tu, P. (2007). Shape and appearance context modeling. *International Conference on Computer Vision*, 1–8.

Welch, G., & Bishop, G. (2006, July). An introduction to the Kalman Filter. *Tech. Rep*.

Wilson, K., Checka, N., Demirdjian, D., & Darrell, T. (2001). Audio-video array source separation for perceptual user interfaces. *Proceedings of Workshop on Perceptive User Interfaces*.

Wu, Y.-S., Chang, Y.-S., Juang, T.-Y., & Yen, J.-S. (2012). An Architecture for Video Surveillance Service based on P2P and Cloud Computing. *2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic.*

Yang, L., & Jin, R. (2006). Distance Metric Learning: A Comprehensive Survey. *Tech. Rep*.

Zhang, R., & Ding, J. (2012). Object Tracking and Detecting Based on Adaptive Background Subtraction. *International Workshop on formation and Electronics Engineering*, 1351-1355.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognitions literature survey. *ACM Computing Surveys*, *35*(4), 399–458.

Zheng, W., Gong, S., & Xiang, T. (2009). Associating groups of people. *Proceedings of the British Machive Vision Conference*, 23.1–23.11.

Zhu, S., & Yuille, A. (1996). Region competition: unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Trans. Pattern Analysis Machine Intelligence, 18*(9), 884–900.

Zotkin, D., Duraiswami, R., & Davis, L. (2002, November). Joint audio-visual trackingusing particle filters. *EURASIP J. Appl. Signal Process*, 1154–1164.

Zou, X., & B. Bhanu. (2001). Tracking humans using multi-modal fusion. *Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition*.

Zou, X., & Bhanu, B. (2005). Pixels that sound. *Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition*, 88–95.