Computers and Electrical Engineering 70 (2018) 736–753

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Video surveillance systems-current status and future trends☆

Vassilios Tsakanikas*, Tasos Dagiuklas

SuiTE Research Group Division of Computer Science, London South Bank University, 103 Borough Road, London SE1 0AA, UK

ARTICLE INFO

Article history:
Received 9 May 2017
Revised 8 November 2017
Accepted 8 November 2017
Available online 14 November 2017

Keywords:
Surveillance
Computer vision
Image analysis
Analytics
Image features
Surveillance analytics
Cloud Computing
Fog Computing
Edge Computing

ABSTRACT

Within this survey an attempt is made to document the present status of video surveillance systems. 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. The most common modalities utilized by surveillance systems are discussed, putting emphasis on video, in terms of available resolutions and new imaging approaches, like High Dynamic Range video. The most important features and analytics are presented, along with the most common approaches for image / video quality enhancement. Distributed computational infrastructures are discussed (Cloud, Fog and Edge Computing), describing the advantages and disadvantages of each approach. The most important deep learning algorithms are presented, along with the smart analytics that they utilize. Augmented reality and the role it can play to a surveillance system is reported, just before discussing the challenges and the future trends of surveillance.

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 image and video analysis algorithms from research areas 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.

Literature suggests that surveillance systems have technically evolved under three generations. The 1st generation (1G) is dated back in 1960s, when analog Close Circuit TV (CCTV) systems were first introduced, mainly for indoor surveillance applications. For that time, 1G systems performed rather satisfying, gaining the trust of the market with deployments in banks, supermarkets, garages, etc. Yet, analogue technology constrained their capabilities, especially for recording and distributing processes. In 1980s, digital imaging evolved surveillance systems to the 2nd generation (2G), offering two major advances. First, compression and distribution have now become more efficient and more cost-effective. Second, computer vision algorithms have been introduced to surveillance systems, offering semi-automated functionalities, such as object tracking and event alerting. Finally, since the early 2000s, we can speak of the 3rd generation of surveillance systems, where fully automated wide-area surveillance systems are explored, aiming to offer reasoning frameworks and behavioral analysis

☆ Reviews processed and recommended for publication to the Editor Dr. E. Cabal-Yepez.
* Corresponding author.
E-mail addresses: tsakaniv@lsbu.ac.uk (V. Tsakanikas), tdagiuklas@lsbu.ac.uk (T. Dagiuklas).

https://doi.org/10.1016/j.compeleceng.2017.11.011
0045-7906/

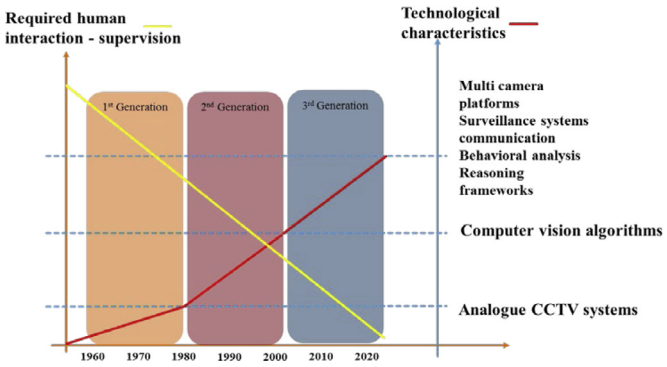

Fig. 1. Evolution of surveillance systems.

functionalities, incorporating and integrating at the same time multi-sensor platforms and data fusion techniques. In Fig. 1, a timeline diagram of the evolution of surveillance systems is depicted.

There are many flavors of Video Surveillance Systems, each one trying to fulfill a piece of the market. 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 offers, 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 behaviors 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. Within the context of surveillance systems, efficiency improvement has two directions. First, the improvement of the video processing algorithms along with the derived video analytics will increase the validity and the accuracy of a surveillance system and second the integration of surveillance systems with cloud infrastructures is expected to improve reliability (e.g. generate alarms under poor lighting conditions etc.), reduce the maintenance costs and increase the response time of the systems.

Surveillance systems have to cope with several challenges, including, but not limited to, algorithmic and infrastructure challenges. Thus, surveillance systems have to adapt with the emerging network and infrastructure technologies, such as cloud systems, 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 scope of this paper is to survey the current status of Video Surveillance Systems, aiming to identify the best practices for image and video processing and analysis and highlights research challenges for next generated systems. Additionally, the applicability of proposed algorithms and architectures will be assessed, in terms of time response and scenarios variety. The paper is structured as follows: Within Section 2, the available video sensors from different surveillance systems are presented, Section 3 describes the different modalities that are commonly used in surveillance systems. In Sections 4 and 5 several approaches for the most studied image and video processing algorithms are analyzed, focusing on video analytics and quality enhancement respectively. In Section 6, computing architectures for boosting the performance of a surveillance system are discussed while in Section 7 the future trends on surveillance systems are drawn.

2. Video sensors

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 is analog video sensors which are used to 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. In Table 1, common formats of analog cameras are provided, along with their resolution are presented. 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. Regarding the FPS (Frames per Second), specification of

Table 1
Resolutions of common analog video cameras.

Analog video format	Resolution
1,080p resolution	1920×1080
720p resolution	1280×720
D1 resolution	704×480 (NTSC for the United States) 720×576 (PAL for Europe)
CIF resolution	352×240
QCIF resolution	176×120

analog video sensors, it can vary between 30 FPS, 15 FPS, 7.5 FPS, 5 FPS and 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.

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 advance video codecs. Consequentially, these sensors can interface directly with network infrastructures and transmit their data over switches and routers. 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-based cameras, which nowadays belong to the HD (High Definition) category can capture video on 1920×1080 resolution and 30 FPS and downgrade to 1280×720 or D1 for 15 FPS. Ultra HD (UHD) video sensors have been also introduced to surveillance systems, pushing the available resolutions to 4 K (3840×2160, usually under15 FPS) or 2048×1536 under 30 FPS.

Finally, since the beginning of 2010, a new type of video cameras has been introduced, the High Dynamic Range (HDR) video sensors. These sensors, which usually operate 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. Additionally, the research community, during the last few years has proposed the usage of High Efficiency Video Coding (HEVC) as an appropriate video coding standard for HDR content. Recently, several organizations including the Blu-ray Disc Association (BDA), the High-Definition Multimedia Interface (HDMI) Forum, and the Ultra-High Definition (UHD) Alliance have decided to adopt a delivery format based on HEVC Main 10, commonly referred to as “HDR10”, for the compression and delivery HDR content.

3. Acquired modalities

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. This section, presents the state of the art in surveillance system with respect to different modalities.

3.1. 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. 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 [1], moving objects are located calculating the sound time delays among the microphones. Applications using sound as modality are multi-object 3D tracking and walking person detection. These 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 in no longer available. Hence, the most common approach for audio-visual fusion is Canonical Correlation Analysis (CCA), using as variables spectral bands for sound and image pixels for video. 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 [2], in which a presumed sparsity of the audio-visual events is exploited. Other approaches for audio-video correlation are proposed in literature. According to these approaches, two groups of multi-variate variables are correlated using the MMI (Maximization Mutual Information) method, while Markov chains are proposed and the audio-video joint densities are estimated using a group of training sequences.

3.2. GPS

Video surveillance systems started to incorporate GPS data when they stopped using fixed cameras and started to incorporate moving cameras. This required addition of an extra layer of meta-data 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 [3], 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 re-projected 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. 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.

3.3. Video

Undoubtedly, video streams are the primarily modality when it comes to surveillance systems. 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.

3.4. 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 have been proposed. 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. 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). Finally, sensor fusion and thermal–visible video registration techniques are proposed in [4], where sensor fusion uses aligned images to compute sum-rule silhouettes, and then constructs thermal-visible object models.

4. Knowledge extraction algorithms

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

4.1. Face detection

Detecting faces within a scene is a mature problem in the area of computer vision. This is because face detection is one of the most widely used processes within surveillance systems, as it is required by many applications such face recognition, face tracking, face analysis for behavioral knowledge extraction [5]. Yet, new applications constantly emerge, such as Human Computer Interaction (HCI), which demands more robust and accurate solutions.

The aim of face detection is to firstly to determine whether any faces are depicted in a scene and secondly to calculate and return the coordinates of the detected faces. This task involves many non-trivial conditions, such as variations in scale, location, orientation and pose, as well as lighting conditions, facial expressions and occlusions.

One common classification of face detection approaches is reported in [6], where four categories are described: (a) template matching methods, where pre-stored face templates are used to decide whether an image contains a face or not, (b) knowledge-based methods, where well established pre-defined rules are used, (c) feature invariant approaches, where structural face features are utilized and (d) appearance-based methods, where models are trained against annotated face data. Nonetheless, [5] suggests that the innovative work of Viola and Jones ([7]) has changed the way modern approaches for face detection are classified, and suggests that face detection algorithms should be categorized to algorithms that are based on rigid-templates and to algorithms that deploy Deformable Parts-based Model in order to model potential deformations among facial points.

One of the most important representatives of the rigid-templates category is the work reported in [7]. Within this work, Viola and Jones proposed a face detector which is based on the integral image, classifier learning with AdaBoost and the attentional cascade structure. Following this concept, new image features have been proposed in order to improve the accuracy of the algorithms. Such features are joint Haar features, which are based on the co-occurrence of multiple Haar-like features and Classification and Regression Tree (CART) based weak classifiers. Another common feature for face detection is based on regional statistics such as histograms, with Histogram of Oriented Gradients (HOG) being the most popular one. Lately, an approach that uses the so-called Integral Channel Features (ICF) with boosting achieved state-of-the-art performance in face detection under various conditions. Regarding the classification schemes, neural networks are widely used, like constrained generative model (an auto-associative, fully connected multilayer perceptron with three large layers of weight) and convolutional neural network (CNN) based approaches.

As far as Deformable Parts-Models (also known as pictorial structures modeling) is concerned, they constitute one of the standard choices for developing generic object detectors. While simple models have been proposed, more complex approaches have provided robust solutions.

4.2. 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 makes face detection a prerequisite process for face recognition. This problem troubles researchers for more than forty years, trying to produce robust, accurate and real-time solutions. The first approaches documented 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.

Nowadays, one can classify the methods for face recognition in three categories; namely holistic matching methods, feature-based methods and hybrid methods. Holistic methods suggest that the whole face region is compared against a face database using specific techniques such as Eigenfaces, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). 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 need to reformulate in order to produce accurate results when the aforementioned features are not visible in the scene. In order to tackle this problem, feature estimation methods have been proposed, mainly taking advantage of face structural constrains. For example, Ahonen et al. [8] proposed a novel approach for face recognition which incorporates both texture and shape information to represent faces. A face is first divided into small blocks from which the Local Binary Pattern (LBP) features are extracted and united into a single feature histogram, which represents the face. In a more recent approach, Lenc and Král [9] propose to use dynamic positions and number of the fiducial points produced by LBP features. The selection of the chosen fiducial points is performed fully automatically, utilizing a set of Gabor filters. Local extrema in the filter responses are detected and used as the feature points. The number of points is further reduced using the K-means clustering algorithm. Finally, 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 [10], 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.

4.3. 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 (e.g. the scene of Fig. 2).[1] Within such setups, a surveillance system should 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.

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 either by 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 tried to address the ID re-identification problem. Some extensive tries can be found in literature. According to most of the published research studies, the problem of ID re-identification

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

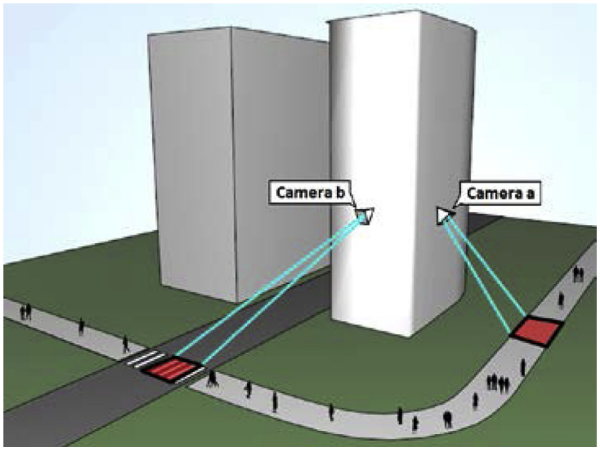

Fig. 2. Person re-identification scenario.

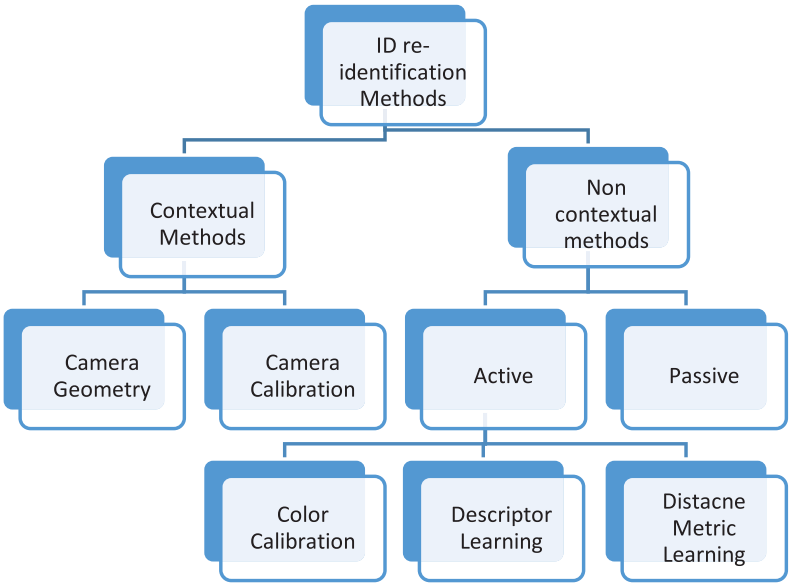

Fig. 3. ID re-identification methods categorization.

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, it is expected 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. The methods and the techniques for ID re-identification are categorized in several methods, as depicted in Fig. 3.

4.3.1. Contextual methods

Contextual methods take advantage of external information such as camera geometry and camera calibration. For example, camera geometry setup is taken into account in order establish intra-camera relationship and increase constrains among the cameras.

Camera geometry is usually determined by correlating activities among cameras with disjoint field of views 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. Then, the assessment of the time delayed motion correlations is used for person re-identification and both temporal and spatial 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, in [11], individuals' height is calculated using homography, to estimate a 3D model. A Panoramic Appearance Map (PAM), uses information from multiple cameras that view the object to produce a single object signature. Other important works in this category are reported in [12]. Hu 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. A modeling approach is also proposed in the literature, where 3D information is extracted from multiple cameras. The proposed model is a 3D Marked 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.

4.3.2. Non-contextual methods

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 modeling is proposed in the literature, 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 has been used, 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 supervized or unsupervized. The supervized approaches require a set of annotated training data, in order to "learn" to detect the desirable features (e.g. person's silhouette), while the unsupervized algorithms utilize clustering techniques in order to estimate different image features (without use of training data). One can categorize these 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.

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 in order to produce descriptive lists of features that better represent an individual's appearance using a bag-of-features approach. Within such approaches, 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.

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.

4.4. 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. The most common techniques to achieve background modeling include median and mead filters.

Temporal Differencing algorithms calculate the difference (on pixel level) between successive video frames, in order to detect the moving object. 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). 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.

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

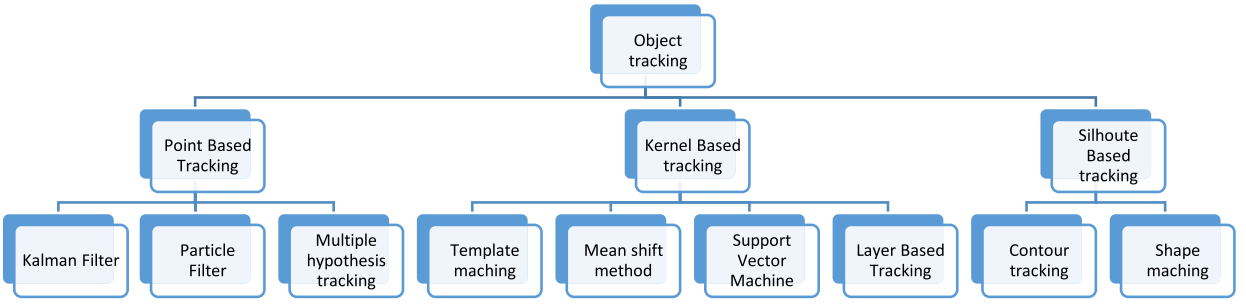

Fig. 4. Object tracking methods.

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. Object tracking can be classified as kernel based tracking, point based tracking and silhouette based tracking [13] (Fig. 4).

The most common point-based approaches utilize Kalman and Particle filters. Kalman filter 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 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 algorithms to track multiple objects.

In Kernel based tracking, kernel denotes 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. Kernel based approaches can be classified in four categories. Template matching algorithms employ a brute force method for 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 are classified as either "tracking object" or "non-tracking object". 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, we discuss about the Silhouette Based Tracking approaches. 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, 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.

5. Quality enhancement algorithms

The knowledge extraction algorithms discussed in the previous section use as input either frames or video streams. Such input is required 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.

5.1. 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, if uniform foreground distribution is assumed, as the intensity of a foreground pixel can randomly take any value

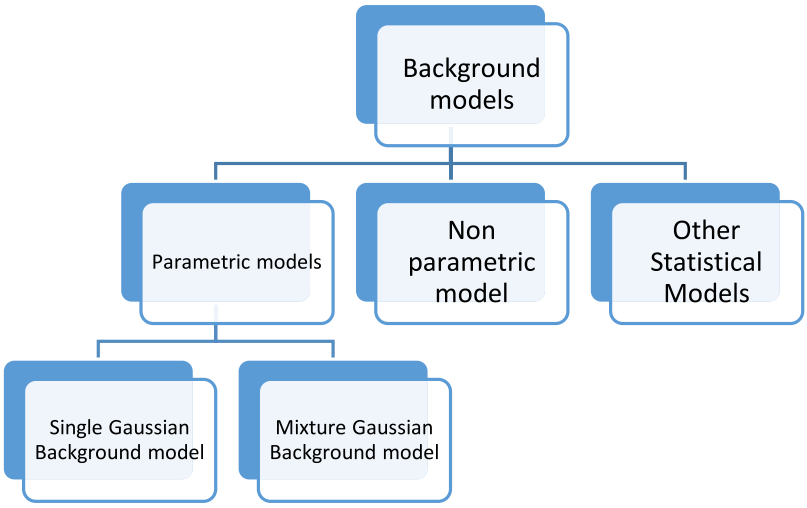

Fig. 5. Background modeling approaches.

(unless specific information about the foreground is available) [14]. 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 to Fig. 5.

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, which was widely reported in literature in the 90s. In 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 in the literature.

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

Lastly, in the literature there have been proposed other statistical techniques for background modeling. For example, linear prediction using the Wiener filter is used to predict pixel intensity given a recent history of pixel values while linear prediction using the Kalman Filter was used in during the previous decades in literature. Another approach has used 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. 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), where a Kalman filter was used in order to update the model. Finally, a biologically-inspired non-parametric background subtraction approach has been proposed, 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 modeling is inspired by the need to have a brightness invariant representation of the scene. Another feature used is optical flow, 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.

5.2. Image/video quality enhancement algorithms

Image/video enhancement algorithms are mandatory for any surveillance system. Low quality sensors and multivariate environmental conditions (e.g. fog, rain, extreme sunshine etc.) 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. 6.

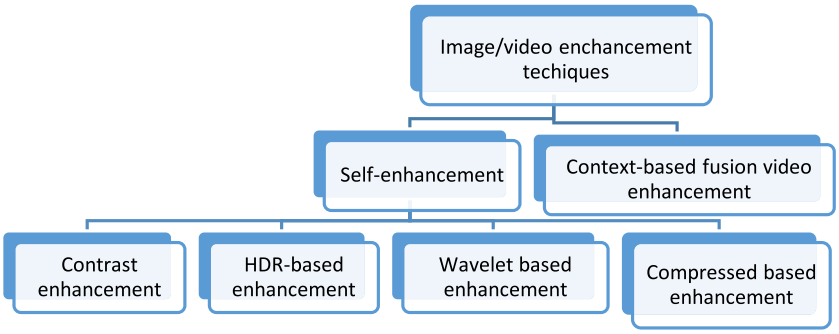

Fig. 6. Categories of Image/video enhancement techniques.

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. Histogram equalization aims to uniformly distribute an image's histogram utilizing density functions. 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. HDR-based enhancement techniques are the second category of self-enhancement methods. 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 or by using image processing algorithms. The third category utilizes wavelet transformation, producing a wavelet image, suitable for processing the image/video. The wavelet techniques utilize wavelet coefficients, wavelet shrinkage denoising or the dual-tree complex wavelet transform. 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.

5.3. 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 [15], 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.

Taking under consideration that nowadays the majority of the installed surveillance systems are CCTV based, there is a great need of addressing issues like robustness to environmental conditions, practical or even automatic effective calibration procedures (applicable to systems of hundreds and even thousands of cameras), dealing with crowded conditions and being able to handle pan-tilt zoom (PTZ) cameras easily. Additionally, most surveillance systems face technical limitations. The most usual ones are improper viewing angle, blind spots in coverage area, improper lighting conditions, improper recording resolution settings and too few cameras with too wide field-of-views. Yet, the suggested techniques (along with the required infrastructures) which are proposed (both from research and industry arena) address these limitations entail costs that can be prohibitive in many applications.

While CCTV systems handle the recorded video stream in house, providing a level of privacy and security on the content, the scene is completely different when it comes to surveillance systems with IP cameras supported by cloud services. Streaming video content over Internet raises security and privacy issues which are difficult to tackle with existing technologies like VPNs or cryptography [16]. The majority of the proposed systems do not deal with these issues which are crucial for a surveillance system, especially if the captured video streams can be viewed as potential law evidence to a court. Thus, there is a great need of designing approaches which will be more robust, more reliable and more secure, increasing the applicability and therefor the economy scale of surveillance systems.

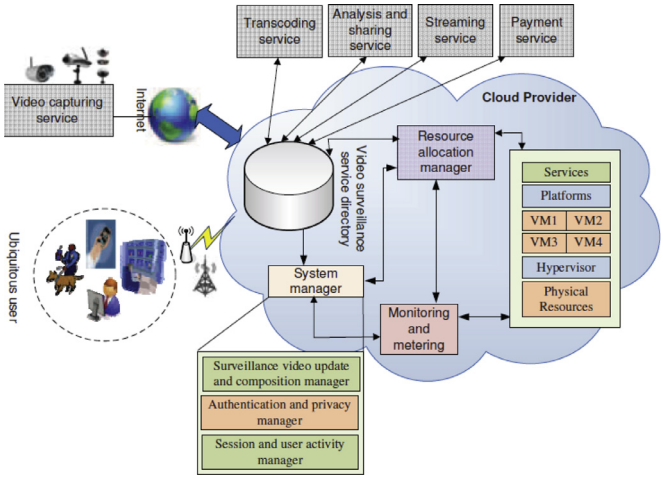

Fig. 7. Proposed conceptual cloud architecture [19].

6. Computing infrastructures

6.1. Cloud-based accelerators

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 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. 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.

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 reported in [17], where the proposed cloud infrastructure is used as SaaS and focus mainly on storage issues, using Amazon S3 platform. On the same track, [18] 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 [19], where VM (Virtual Machines) resources are tuned based on QoS requirements, as depicted in Fig. 7.

6.2. Fog/edge based accelerators

As discussed in the Section 6.1, the concept of introducing cloud infrastructures and services into surveillance systems resolves (partially or fully) major limitations of current systems, such as lack of computational power and restrictions in storage capacity. Yet, this approach introduces some new challenges that need to be addressed in order to end up with a surveillance system that will be capable of meeting the requirements of the end users. More specifically, a surveillance system that utilizes cloud infrastructures needs to take into account the latency and the extra communication cost that is introduced between the sensors and the cloud infrastructure. Sending video streams to the cloud is by no means cost effective, especially if the video sensor has large resolution (e.g. HD video), while at the same time bearing in mind the "best-effort" character of IP networks, the latency that is introduced is not only large but also fluctuating. These network characteristics prevent a cloud surveillance system from "reacting" to (near) real time events, such as car accidents.

A solution to these issues could be Fog Computing. Fog Computing (Fig. 8) is a paradigm that extends Cloud Computing and services to the edge of the network. Thus, one should not face Fog Computing as a competitor to Cloud Computing but a complement technology that improves the characteristics of Cloud services. One can describe Fog Computing as a distributed computation concept which is installed near (in terms of communication latency) the production of raw data. Fog Computing

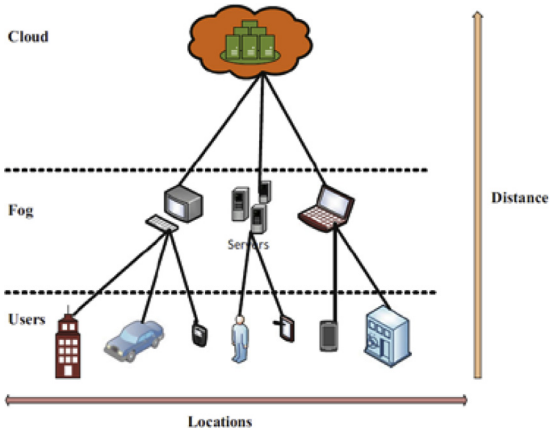

Fig. 8. The Fog Computing conceptual architecture [20].

comprises many (and sometimes heterogeneous) devices that are capable of communicating and reallocating computational tasks. These devices have enough power to perform non trivial computational tasks, but in no case they match the capacity of a cloud system. The main advantage of Fog Computing is that if can offer to a surveillance system a first level of analytics extraction and decision making with minimum network traffic overhead and latency.

The research community during the last few years tries to utilize the concept of Fog Computing to surveillance systems, where video streams from Google Glasses© where captured and processed by either Google Glass device or the user's mobile devices (e.g. smartphone), depending on the battery life of the devices and the required computational power. The architecture is tested on several video processing tasks, such as face recognition and Optical Character Recognition (OCR). An urban traffic management and car accident system is described and tested in [20], where Fog Computing enables the near-real time vehicle tracking and its speed calculation.

Following and projecting the logic of the last two sections, Edge Computing is introduced to surveillance systems. As defined in [21], Edge Computing refers to the enabling technologies allowing computation to be performed at the edge of the network, on downstream data on behalf of cloud services and upstream data on behalf of IoT services. Within the context of surveillance systems, Edge Computing refers to transferring computational and storage capacities from datacenters to the video sensors (or any other kind of utilized sensors), minimizing further (comparing with Fog Computing) and eliminating network latency. The paradigm of Edge Computing deployed to a surveillance system will require the usage of special hardware and / or software aside each video sensor. This hardware/software will be able to perform a first level of video processing which can boost the performance of the surveillance system, in terms of response time and communication costs. For example, the edge systems could calculate the features (e.g. HOG descriptors) from the captured video stream and forward them to the next tier, decreasing the network requirements, as performed in [22].

In another research study, body-worn cameras are suggested to be used by police officers, aiming to provide specific analytics for law enforcement [23], while in [21] case-studies for cloud offloading and video analytics at the edge of a surveillance system are explored (Fig. 9).

Taking under consideration the described Cloud, Fog and Edge Computing concepts within the context of surveillance systems, a promising approach is a "blended" architecture (Fig. 10), where certain characteristics (e.g. low-cost and proximity of the Edge layer, connectivity and power sustainability of the Fog layer and computational power and storage of the Cloud layer) of each approach are utilized, in order to maximize the efficiency of the system.

6.3. Deep learning methodologies

Deep learning approaches are intensively utilized during the last decade for addressing some of the most challenging problems regarding visual content, such as image classification, knowledge extraction and object identification. While the concept of deep learning regarding visual context was initially introduced through many years before, inspired by a biological model of the cat's visual system, it only produced tangible implementations at the end of the previous decade. One can mention three key developments. Namely, the high increase in the computational power and in the capacities of the processing hardware, the exponential decrease of the hardware's cost and the substantial advances in the machine learning algorithms.

Deep learning algorithms are a subclass of machine learning algorithms, which have the capacity of discovering multiple levels of distributed representations. The key word is "discovering", which implies that deep learning algorithms can identify the most important features that should be used for performing an information representation, such as object identification or human pose estimation. In order to achieve this, deep-learning approaches usually requires a (very) large dataset of annotated data. The features that deep learning approaches retrieve usually have a very important characteristic, when it

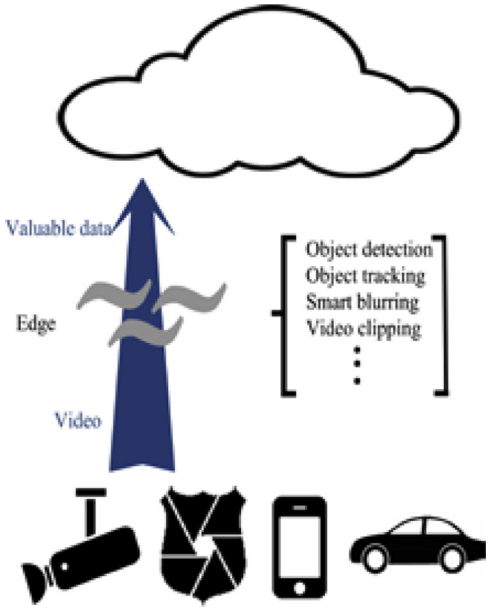

Fig. 9. Overview of Edge Computing concept [22].

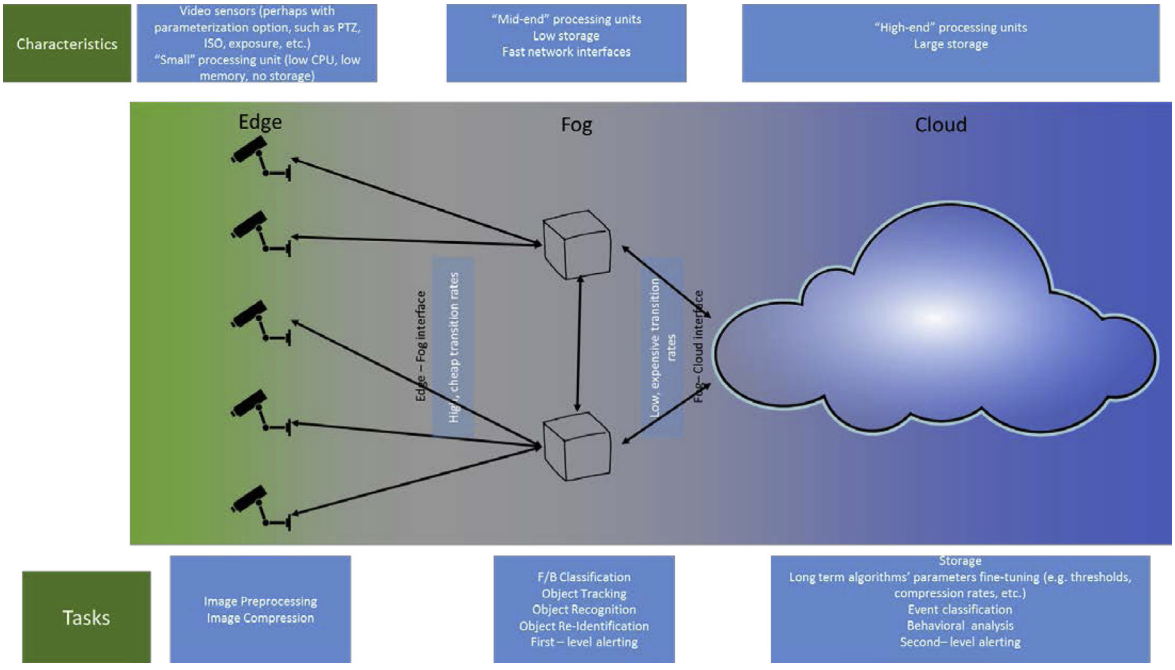

Fig. 10. Cloud, Fog & Edge blended conceptual architecture.

comes to visual content analysis. They are invariant to irrelevant variations of the input. While for humans this task is trivial (e.g. identifying a lion regardless its pose), for many image processing algorithms, a change of an object's pose can alter the labeling output.

While the research community has proposed many algorithms, techniques and methodologies for deep learning algorithms, one can categorize the deep learning approaches in four classes [24]. Namely, (i) Convolutional Neural Networks (CNNs), (ii) Restricted Boltzmann Machines (RBMs), (iii) Autoencoder-based methods and (iv) Sparse Coding-based methods.

CNNs is probably the most common deep learning approach in the visual context. CNNs utilize multiple layers, which are trained in a robust manner. Effectiveness and robustness of CNNs have been proved by many research works in var-

ious computer vision applications. Some of the most important works in this directions are AlexNet ©, Clarifai©, VGG©, GoogLeNet© and SPP©.

Restricted Boltzmann Machines (RBMs) were originally introduced by [25]. An RBM, which is basically a generative stochastic neural network, is a modified Boltzmann Machine, with the constraint that the visible units and hidden units must form a bipartite graph. This constraint allows for more effective training algorithms, such as the gradient-based contrastive divergence algorithm. Some of the most representative research works, which utilize RBMs are Deep Belief Networks, Deep Boltzmann Machines and Deep Energy Models.

Auto encoders is a special class of artificial neural networks, which utilize an unsupervized learning algorithm that applies backpropagation, setting the target values to be equal to the input values. Auto encoders are trained to reconstruct their own inputs, which are then used for the training phase (this explains the auto in their name). In other words, they are trying to learn an approximation of a function, so as to produce an output that is similar to the input data. This results to the output vectors having the same dimensions as the input vectors. The encoder brings the data from a high dimensional input to a "bottleneck layer", where the number of neurons is smaller than the input and output layers. Then, the decoder takes this encoded input and converts it back to the original input image. The latent space is the space in which the data lies in the bottleneck layer. The latent space contains a compressed representation of the image, which is the only information the decoder is allowed to use to try to reconstruct the input as faithfully as possible. To perform well, the network has to learn to extract the most relevant features in the bottleneck. Some of the most important applications that utilized auto encoders are Sparse Autoencoder, Denoising Autoencoder and Contractive Autoencoder.

Sparse coding aims to learn an over-complete set of basic functions in order to describe the input data. Two of the most important advantages of the sparse coding are (i) it can reconstruct the descriptors by using multiple bases and capturing the correlations between similar descriptors which share bases and (ii) the sparsity allows the representation to capture salient properties of images and videos. The most important research works that utilize sparse coding are Sparse Coding SPM, Laplacian Sparse Coding, Local Coordinate Coding and Super-Vector Coding.

7. Future trends

7.1. Surveillance systems and augmented reality

Augmented Reality, in the context of surveillance systems, refers to the information depicted on the operator's screen(s) on top of the video stream captured by the surveillance cameras of the system. 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 [26] a scheme is proposed for observing multiple video streams and a visualization system is proposed by merging dynamic imagery with geometry model of a battlefield visualization. In [27], 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 several research studies.

On a similar context, within [28], a surveillance and rescue system is described which automatically combine computer-generated imagery with real-world imagery in a portable electronic device by retrieving, manipulating, and sharing relevant stored videos. Proposing similar technologies, the work presented in [29] describes a visualization system for video surveillance based on an Augmented Virtual Environment (AVE) that fuses dynamic imagery with 360 and 3D models in a real-time display to assist observers and users to easily and effectively comprehend multiple video streams of temporal data and imagery from random views of the scene, where moving objects are detected and tracked in video streams and visualized as 3D elements in the AVE scene display in real-time. Finally, numerous research works have been presented where augmented reality is used to support operators watching video streams from surveillance cameras by offering functionalities like removing immobile items (over certain time frame) from a scene, providing text-based and sound-based messages or even proposing areas in the scene that the operator should pay attention to.

7.2. Future trends

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. Yet, a lot of work still needs to be done. 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

Fig. 11. Research trends of surveillance systems.

systems, researchers working on video analytics should bring surveillance to the next level, working on the following topics (Fig. 11).

7.2.1. Cloud/Fog/Edge 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 must be designed, which will maintain the accuracy of the video analytics algorithms while reducing the aforementioned costs. Cloud technology, as argued in paragraphs 6.2 and 6.3, can be supported and extended by Fog computing and Edge computing. More specifically, Fog and Edge computing can address the delay overhead that cloud services usually impose to a surveillance system by transferring computational power closer to the source of the event. By calculating features and analytics close to the sensors decreases the required network bandwidth and increases the response time of the system. Thus, these approaches can be used to cutting-edge approaches like automatic drone navigation or automatic field of view rearranging.

7.2.2. 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.

7.2.3. Modality fusion

Apart from video, which is the dominating sensing technology for surveillance systems, other modalities can facilitate monitoring and alerting tasks. Such modalities are 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 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.

7.2.4. 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.

7.2.5. 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 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.

7.2.6. 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.

7.2.7. Virtual reality

As the number of the video sensors of a scene/area increases, the operator's monitoring work becomes non-trivial, as she/he has to constantly pay attention to multiple screens. As already argued, augmented reality can facilitate this workflow through adding an intelligent layer to the monitoring screens. Another approach to achieve this workflow facilitation is Virtual Reality (VR), where a set of algorithms and techniques will reconstruct a 3D (360°) world from the video sensors, in which the operator will be able to walk through and observe certain features, such as objects and individuals. While such solutions have been proposed in the literature ([30]) a lot of work still needs to be done, improving the required algorithms for gaze direction computation, camera scheduling, collaborative tracking and Virtual Reality content streaming.

7.2.8. Deep learning algorithms

While supervized deep learning algorithms have performed extremely well, when compared with other approaches, unsupervized learning is expected to play an important role in reviving interest in deep learning. Unsupervized learning is expected to produce representations and relationships which are not obvious even to animals and humans. Also, deep learning approaches are expected to mimic even further human vision, producing systems that are trained end-to end to decide where to in the field of view the system should focus.

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.

8. Conclusion

After surveying the current status of surveillance systems, both from the algorithmic and the systemic / infrastructure point of view, several conclusions can be drawn regarding the available technology nowadays, the technological limitations and the future challenges of the area.

Video surveillance systems have been introduced almost fifty years ago, through CCTV systems, requiring a substantial number of manpower, analog to the number of the installed video sensors, leading to high operational costs. The majority of the research studies on surveillance systems the last five decades are trying, to substitute the operator with a video processing algorithm which will be able to perform certain tasks. While all of the effort put on this non-trivial mission has produced some really innovative and brilliant algorithms, these are only limited to a small set of tasks, like face recognition and object detection and tracking. The accuracy of these algorithms is usually far from satisfying when the scene conditions are not perfect.

Research developments during the last decade have revitalized the expectations for automatic surveillance systems. These developments mainly involve machine-learning, deep-learning algorithms and distributed computational infrastructures, like cloud, Fog and Edge Computing. These methodologies, combined together, are expected to improve the accuracy of surveillance algorithms, proposed new smart analytics and reduce the response time of the systems, in order to produce meaningful alerts.

As already discussed, there are a lot of research challenges that need to be addressed. Among these, special attention needs to be placed on optimization techniques that will automatically redistribute the computational power among edge, fog and cloud agents, based on specific performance, cost and privacy criteria. Such optimization techniques will boost the performance of the surveillance systems, enabling at the same time a new paradigm, Video Surveillance as a Service (VSaaA).

Acknowledgments

This paper is dedicated to the memory of Themistoklis Nikolopoulos, who suddenly passed away on May 2017.

References

[1] Zotkin D, Duraiswami R, Davis L. Joint audio-visual tracking using particle filters. EURASIP J Appl Signal Process November 2002:1154–64.
[2] Zou X, Bhanu B. Pixels that sound. CVPR'05: proc. 2005 conf. computer vision and pattern recognition; 2005.
[3] Kumar R, Sawhney H, Samarasekera S, Hsu S, Tao H, Guo Y, et al. Aerial video surveillance and exploitation. Proc IEEE 2001;89(10):1518–39.
[4] Torabi A, Massé G, Bilodeau G-A. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. Comput Vis Image Understand 2012;116(2):210–21.
[5] Zafeiriou S, Zhang C, Zhang Z. A survey on face detection in the wild: past, present and future. Comput Vis Image Understand September 2015;138:1–24.
[6] Yang HM, Kriegman JD, Ahuja N. Detecting faces in images: a survey. IEEE Trans PAMI 2002;24(1):34–58.
[7] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Computer vision and pattern recognition (CVPR 2001) Kauai, HI, USA; 2001.
[8] Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns. European conference on computer vision - ECCV 2004 Prague; 2004.
[9] Ladislav L, Král P. Local binary pattern based face recognition with automatically detected fiducial points. Integr Comput-Aided Eng 2016;23(2):129–39.
[10] Kamgar-Parsi B, Lawson W, Kamgar-Parsi B. Toward development of a face recognition system for watchlist surveillance. IEEE Trans Pattern Anal Mach Intell 2011;33(10):1925–37.
[11] Lantagne M, Parizeau M, Bergevin R. Vip: vision tool for comparing images of people. 16th international conference on vision interface; 2003.
[12] Baltieri D, Utasi A, Vezzani R, Csaba B, Szirányi T, Cucchiara R. Multi-view people surveillance using 3D information. Eleventh international workshop on visual surveillance Barcelona, Spain; 2011.
[13] Athanesious JJ, Suresh P. Systematic survey on object tracking methods in video. Int J Adv Res Comput Eng Technol 2012;1(8):242–7.
[14] Elgammal A. Background subtraction: theory and practice. In: Asari VK, editor. Wide area surveillance. New Brunswick, Berlin: Springer-Verlag; 2014. p. 1–21.
[15] Porikli F, Brémond F, Dockstader SL, Ferryman J, Hoogs A, Lovell BC, et al. Video surveillance: past, present, and now the future. IEEE Signal Process Mag 2013;30(3):190–9.
[16] Costa DG, Figuerêdo S, Oliveira G. Cryptography in wireless multimedia sensor networks: a survey and research directions. Cryptography 2017;1(1).
[17] Rodríguez-Silva DA, Adkinson-Orellana L, González-Castaño FJ, Armiño-Franco I, González-Martínez D. Video surveillance based on cloud storage. IEEE fifth international conference on cloud computing Honolulu, HI, USA; 2012.
[18] Li Q, Zhang T, Yu Y. Using cloud computing to process intensive floating car data for urban traffic surveillance. Int J Geograph Inf Sci 2011;25(8):1303–22.
[19] Hossain MS, Mehedi Hassan M, Al Qurishi M, Alghamdi A. Resource allocation for service composition in cloud-based video surveillance platform. IEEE international conference on multimedia and expo workshops Melbourne, Australia; 2012.
[20] Chen N, Chen Y, Ye X, Ling H, Song S, Huang C-T. Smart city surveillance in fog computing. Advances in mobile cloud computing and big data in the 5G era, 22. Springer International Publishing; 2017.
[21] Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: vision and challenges. IEEE Internet Things J 2016;3(5):637–46.
[22] Zhang Q, Yu Z, Shi W. Demo abstract: evaps: edge video analysis for public safety. IEEE/ACM symposium on edge computing (SEC) Washington, DC, USA; 2016.
[23] Corso JC, Alahi A, Grauman K, Hager GD, Morency L, Sawhney H, et al. Video analysis for body-worn cameras in law enforcement A white paper prepared for the Computing Community Consortium committee of the Computing Research Association; 2015.
[24] Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding. Neurocomputing 2016;187:27–48 C.
[25] Hinton G, Sejnowski TJ. Learning and relearning in boltzmann machines. In: Parallel distributed processing: explorations in the microstructure of cognition, 1. Cambridge, MA, USA: MIT Press; 1986. p. 282–317.
[26] Hall B, Trivedi M. A novel graphical interface and context aware map for incident detection and monitoring. 9th world congress on intelligent transport systems Chicago, Illinois, USA; 2002.
[27] Kumar R, Sawhney H, Guo Y, Hsu S, Samarasekera. 3D manipulation of motion imagery. International conference on image processing Vancouver, BC, Canada; 2000.
[28] S. Adhikari, T. Dunn and E. Hsiao, Augmented reality system for product identification and promotion. US Patent US20120113145 A1, 2012.
[29] Ismail SO, Hu J, You S, Neumann U. 3D video surveillance with augmented virtual environments. First ACM SIGMM international workshop on Video surveillance (IWVS 2003) Berkeley, California; 2003.
[30] Du R, Bista S, Varshney A. Video fields: fusing multiple surveillance videos into a dynamic virtual environment. 21st international conference on web3d technology (Web3D '16) Anaheim, California; 2016.

Vassilios D. Tsakanikas received his Diploma Degree in Electrical and Computer Engineering from NTUA and his M.Sc. in Computer Science from AUEB. His research interests are signal processing, computer vision and machine learning. Vasilis is a Ph.D. student at the London South Bank University and a member of the SuITE research group (https://www.lsbu.ac.uk/research/centres-groups/sites/smart-internet-technologies-suite).

Tasos Dagiuklas is a leading researcher and expert in the fields of Internet and multimedia technologies for smart cities, ambient assisted living, healthcare and smart agriculture. He is the leader of the newly established SuITE research group (https://www.lsbu.ac.uk/research/centres-groups/sites/smart-internet-technologies-suite) at the London South Bank University where he also acts as the Head of Division in Computer Science.