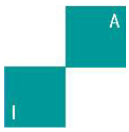
Inteligencia Artificial 20(59) (2017), 53-69
doi: 10.4114/intartif.vol20iss59pp53-69



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

A Review on Intelligent Monitoring and Activity Interpretation

José Carlos Castillo

Social Robots Group, Universidad Carlos III de Madrid, 28911-Leganés, Spain
jocastil@ing.uc3m.es

Antonio Fernández-Caballero

Instituto de Investigación en Informática de Albacete, Universidad de Castilla-La Mancha, 02071-Albacete, Spain
Antonio.Fdez@uclm.es

María T. López

Instituto de Investigación en Informática de Albacete, Universidad de Castilla-La Mancha, 02071-Albacete, Spain
maria.lbonal@uclm.es

Abstract This survey paper provides a tour of the various monitoring and activity interpretation frameworks found in the literature. The needs of monitoring and interpretation systems are presented in relation to the area where they have been developed or applied. Their evolution is studied to better understand the characteristics of current systems. After this, the main features of monitoring and activity interpretation systems are defined.
Resumen Este trabajo presenta una revisión de los marcos de trabajo para monitorización e interpretación de actividades presentes en la literatura. Dependiendo del área donde dichos marcos se han desarrollado o aplicado, se han identificado diferentes necesidades. Además, para comprender mejor las particularidades de los marcos de trabajo, esta revisión realiza un recorrido por su evolución histórica. Posteriormente, se definirán las principales características de los sistemas de monitorización e interpretación de actividades.

Keywords: Monitoring, surveillance, fusion, activity interpretation, frameworks.
Palabras Clave: Monitorización, vigilancia, fusión, interpretación de actividades, marcos de trabajo.

1. Introduction

New-generation systems for monitoring and activity interpretation are characterized by a significant improvement in the chances of quickly and efficiently transmitting data, voice and video. One of the features required to date is real-time monitoring of the environment under control. In addition, current proposals interpret what happens in the scenario from data provided by sensors to predict the actors' actions. In recent years, interest in these systems has increased greatly, especially following the events of September 11th. A proof of this are the numerous conferences and journals available today on this subject. In the industrial field these systems have also undergone extensive proliferation. On the one hand, there are monitoring systems for transport: we can find applications in airports [83, 84, 86], ports [35, 85, 69], railway and underground stations [54, 1, 14], and vehicle control [49, 81]. On the other hand, there are systems for monitoring public places like banks, shops, homes and garages [48, 22]. In addition, there are monitoring systems for human activities [14, 29, 36]. Also the demands of industry [33] and, of course, the military have been attended.

ISSN: 1988-3064(on-line)

The needs of commercial monitoring systems are different from those developed for academia. Commercial systems call for special-purpose hardware, faster communication networks and intelligent cameras for tasks such as detection of intruders as well as suspicious objects [64, 40]. On the other hand, academic monitoring systems focus on getting algorithms for detection, recognition and tracking of objects and people, as well as recognition of human activities [30, 77]. There is also a growing work in systems composed of multiple heterogeneous sensors, fusing information from different sensors to cover large monitored areas [54, 29, 93, 74]. Within the systems dedicated to monitoring and activity interpretation, there is a large branch in surveillance systems. These systems have been widely studied over the last decades and great effort, both academic and industrial, has been devoted. Therefore, the evolution of these systems is discussed below. It can be affirmed that these developments have served as a model for other monitoring systems.

The remainder of this paper focuses on systems for monitoring and activity interpretation. Firstly, Section 2 provides an overview of the evolution of such systems, from their inception in the 60's and 70's to the most advanced proposals present today. Thus, by observing the historical changes that the systems for monitoring and activity interpretation have suffered, the types of systems that exist today will be clearly explained. After this, some basic concepts are defined to understand the terminology as found in the literature. To this end, Section 3 summarizes the main levels of monitoring systems, collecting some of their main features and proposals found in the literature to help illustrate them. Next, Section 4 introduces the concept of multisensor fusion, essential in today's monitoring systems. After learning the key concepts related to frameworks for monitoring and activity interpretation, Section 5 delves into the study of existing monitoring frameworks, providing classifications for better understanding. Finally, Section 6 provides a synthesized overview and a series of conclusions drawn from the study performed.

2. Evolution of Monitoring and Activity Interpretation Systems

The evolution of surveillance systems can be studied through the devices they use. Starting from initial systems where a camera is connected directly to a monitor where monitoring tasks were manual, we now reach systems capable of autonomously monitoring the real world, identifying the actions that happen in it [80, 32].

2.1. First Generation Systems

Systems belonging to a first generation, introduced in the U.S. and England in the 60s and 70s, are the closed circuit television (CCTV) systems. This type of system consists in a series of cameras distributed in the environment to be monitored, connected to one or more screens. In these systems, the use of analog technologies to implement image storage has several drawbacks such as the cost of maintenance and the dependence on a human operator to detect abnormal situations, which leads to problems due to fatigue caused by constant vigilance. This derives into an attention deficit, thus increasing the possibility of errors. Despite the drawbacks, this type of system is still widespread, particularly in retail and industry. The most advanced CCTV systems incorporate digital video storage and new features that expand their capabilities (e.g. omnidirectional cameras and night vision or modules for motion detection). These systems have improved today thanks to the benefits offered by communication networks, finding digital surveillance systems that send information to different types of components such as computers to process images or digital video recorders (DVR) devoted to storage.

This first generation also introduced the concept of IP-surveillance, which implied an evolution of traditional CCTV systems incorporating digital transmission services over the Internet using IP (Internet Protocol). Systems based on IP-surveillance consist of one or more IP cameras capable of sending recorded footage via the network, which provides advantages over traditional CCTV systems. One advantage is the ability to scan images: (i) ability to perform searches, (ii) greater ease of use, (iii) higher image quality and persistence over time, (iv) ability to record and play simultaneously, (v) compressibility of the information. Another advantage concerns transmitting data over the network: (i) ability to control and remote playback, (ii) remote IP storage, (iii) ease of distribution of images, and (iv) capability to send alerts through the network (via e-mail, for example).

To summarize, we can say that the use of digital technologies has provided advantages over traditional analog technology in the first generation of surveillance systems by simplifying maintenance and storage, the possibility of mining and image processing, as well as the ability to provide the system with portability. The opening lines of research focus on image compression and retrieval of multimedia data in this generation's algorithms.

2.2. Second Generation Systems

The second generation of surveillance systems is born in the $80's$ from the combination of CCTV systems and IP-Surveillance with computer vision algorithms and artificial intelligence techniques [31]. These systems seek to reduce the dependence on human operators seen in the first generation systems, interpreting the different

behaviours that occur in the surveillance area. Today, the interpretation of these behaviours is an area where many research groups are turning their efforts. Although there is currently no consensus about the features that these systems must meet, we can establish a set of common requirements:

- The systems must operate in real-time despite the large amount of information to work with. In addition, under certain conditions, a low response time and performance are imperative (e.g. in case of disasters or accidents).
- The systems must deal with uncertainty in a real environment. Like with human operators, there are certain situations in which it is not completely clear whether an event is being triggered, or even where the event is happening.
- The systems must create and manage knowledge bases since it is necessary to know the real world in order to interpret the actions that occur therein. The system must know both the elements of the scene and the relationships between them. Interestingly, this is related to incorporating algorithms able to perform machine learning based on the results obtained previously, thus improving decision-making processes.

To summarize we note that second generation systems pursue the automatic interpretation of real scenes. With respect to its predecessors, this greatly reduces human work as the system assumes an important part of the screening process. There are several proposals for the generation of more efficient algorithms for recognition of events and activities as well as for learning and decision-making [57, 8].

2.3. Third Generation Systems

Third generation surveillance systems share the progress and processing techniques of the above, adding capabilities for distributed processing [80]. This type of system consists of a large number of sensors of different types that provide real-time information for environment monitoring. The fact that the sensors are distributed in these systems provides a number of added advantages such as robustness, in the event that one or more sensors fail, or the possibility of distributing the computation across multiple nodes of the network, since the processing can be decentralized. Because there might be a large number of sensors connected to the same system using a network it is necessary to use middle ware technologies to provide a set of services capable of enabling distributed applications to run on platforms of different types (heterogeneous). Furthermore, the distribution of high-level information between sensors also poses a problem, as objects may pass through the coverage areas of different sensors, which requires the existence of mechanisms that ensure the consistency of data. Also, the size of data managed by networked surveillance systems grows exponentially with respect to volume, velocity, and dimensionality, due to the widespread use of embedded surveillance cameras and sensors [41].

Although we can say that the three generations use to coexist in a business-oriented environment, it is noteworthy that the academic community efforts are directed towards the development of second and mainly third generation solutions. The proposals framed in the second generation are those where processing techniques are proposed to solve specific problems (tracking objects in a sequence of images, for example). In recent years, proposals that link the processing to the distribution of sensors in a monitored environment are emerging, giving rise to third generation systems. In turn, these systems pose several problems associated with the distribution of information. It is necessary to take into account the integration of heterogeneous communication protocols and data of various kinds and formats. Also, it is necessary to create new design methodologies able to support the capture, processing and distribution of information in a distributed manner. Third generation systems should include mobile sensor networks and communication protocols adapted to these needs. For all these reasons, an expansion of third generation systems is necessary to allow interoperability between processing nodes, regardless of the connected sensors in each of them. Therefore, the limitations of third generation systems should be overcome from what might be called fourth-generation monitoring and activity interpretation systems, where ambient intelligence capacities are included. This implies the ability of systems to sense and respond to human presence, helping them in their daily tasks. Before proceeding to the next section which studies the functionality of systems of second and third generation in depth, a summary showing the main advantages and disadvantages of different generations of surveillance systems is provided in Table 1.

3. Monitoring and Activity Interpretation Levels

Before describing the frameworks and systems for monitoring and activity interpretation it is necessary to define the main features of their levels. If we consider a framework such as a bookcase where each shelf corresponds to one level, the techniques discussed in this section would be boxes that are placed on different shelves, each in a well-defined level which can supply features. Initially, we can define three categories. The first, called *segmentation*, is responsible for splitting images to extract the objects or parts that form it. The second class includes methods

Cuadro 1: Evolution of surveillance systems.

	Techniques	**Advantages**	**Drawbacks**	**Research**
First Generation	Analog CCTV systems	Robust Mature technology	Those related to analog technology in storage and diffusion	Digital vs. analog Digital video storage Video compression
Second Generation	Combination of computer vision techniques with CCTV systems	Improving the efficiency of CCTV systems	Precise detection and tracking algorithms are required for behaviour analysis	Robust and real-time computer vision algorithms behaviour pattern learning Generation of natural language interpretation from statistical analysis
Third Generation	Wide-area automatic surveillance	More precise and robust information through sensor fusion Information distribution	Those related to information distribution (integration and communication) Design methodologies Mobile and multisensor platforms	Centralized vs. distributed processing Information fusion Inclusion of probabilistic models Multicamera surveillance

for *tracking* the objects over time, and generates primitive for the third category, the *interpretation of events and activities*. Logical, temporal and spatial relationships, among others, define the structure of the events and activities. We can identify the levels of segmentation, classification and monitoring as belonging to the category of abstraction or collection of features, while the analysis of activities and behaviour would be included within the event (lower level) and activities (higher level) model.

3.1. Segmentation

Image segmentation is defined as the process by which an image is divided into the parts or objects that constitute it [34]. The objective is the location of significant areas of the image, such as imperfections in a tool, urban areas in the event that you are working on a map, humans, etc. This process involves two major tasks. On the one hand it is necessary to perform the decomposition of the image for further analysis and, second, the image pixels are organized into higher-level units that acquire meaning for later analysis.

According to a known categorization [34], there are two image properties on which segmentation algorithms are based. On the one hand we find the *similarity*. This property means that the pixels belonging to a same object must present a uniform appearance. However, this assumption is subject to factors such as lighting, noise, or reflections projected by the objects. This means that similarity is not always fulfilled in real images. Despite its limitations, this property provides a great power by allowing the use of thresholding techniques. Thresholding techniques are used to highlight areas of the image that satisfy a certain property (usually areas that are within a certain range of brightness), discarding the rest. There are many ways to determine the optimal threshold to be applied to an image or image sequence based on the sought objects of interest, although the most widely used is that of Otsu [66]. This method divides the image pixels into two classes, one of which contains the objects of interest and the other the rest (which is called the background). The goal is to find the optimal threshold to separate these two classes so that the intra-class variance is minimized. While the method is easily expendable to obtain three or more classes, the author considers that the thresholds become too complex and there is a big impact on the loss of power of discrimination. Alternatively, various techniques have been proposed for choosing the thresholds. For example, [78] use fuzzy logic techniques to establish the optimal threshold, combining the results with Canny's algorithm for edge detection [9].

The second property used as basis in segmentation algorithms is known as *discontinuity*. This property is based on looking for areas where a high image contrast is perceived. These areas use to correspond to the boundaries of the objects of interest. There are many edge detection algorithms based on this property that can plot the contours of the objects that form the image [96, 9]. Thus, edge pixels are considered when presenting different grey levels to those of their neighbours. Many times, algorithms based on this property are used to refine the results of methods previously applied in order to better define the objects of interest. One of the leading edge detection algorithms is that of Canny [9]. This algorithm is based on the definition of a set of objectives for the calculation of points belonging to edges. These objectives are chosen according to a number of criteria:

- The algorithm must detect and define as many real edges in the image as possible.
- The edges should be detected as close as possible to the real edge in the image.
- The edges should be detected only once in the image, trying to mitigate the false edges caused by noise.

Segmentation techniques can also be classified based on the methodology used. On the one hand, proposals are based on *statistics and probability*. These techniques add concepts of data mining, incorporating a top layer to update and refine the results. These techniques use the temporal coherence of image sequences, collecting statistical

Cuadro 2: Techniques used in object segmentation.

Methodology	Features	Implementations	Publications
Statistical and probabilistic techniques	Use of the sequences temporal coherence	Snakes Geometric structures based on shapes	[46], [70], [59]
Region growing techniques	Partitioning of the images in different regions with common features	Wavelet transform	[44]
Graph-based techniques	Interpretation of the images as a set of arches and nodes	Minimal recovering graph Graph partitioning	[75]
Learning-based techniques	*Supervised*: Selection of a set of pixels with previous knowledge *Non-supervised*: Estimation of the optimal number of image regions	k-means Pose prior maps	[91], [52] [86]

information from it. There are also suggestions that use artificial intelligence concepts, adding a machine learning layer in such cases. As an example of this family of methods we find a model of active contours called snakes [46] which are modelled by a curve defined by polynomials. The polynomials are given by the gradients of the lines and edges of the image. These snakes have been applied to tasks such as edge detection, corner detection, tracking moving objects and matching stereo images [70]. In this line, we find a work which uses a statistical method to detect humans using geometric structures based on the human body [59]. The representation is based on statistical analysis of the distance between the different parts of the images [55].

Following the proposed classification of segmentation methods, we find those based on *region growing*. As indicated at the beginning of this section, the idea common to all segmentation methods is to partition an image into a set of regions. Region growing techniques partition the image iteratively through successive passes on the grounds that pixels belonging to the same object possess common features, especially in terms of intensity of the grey level. These techniques are an alternative to the use of thresholds for segmentation. An example of such techniques uses the wavelet transform, dividing the image in successive sub-images in the frequency domain so that the former contains more general details of the images, while sub-images contain the most unique and distinctive details [44].

Another family of proposals groups the methods based on *graphs* [26, 21]. These proposals interpret an image as a graph, establishing a set of arcs and nodes based on various techniques such as the minimum cover graph. A concept often found in the literature associated with the techniques of graph is the *cut*. When dividing a graph into two disjoint graphs it is necessary to remove the edges connecting the two graphs. A cut is defined as the weight of the edges that had to be removed, with the optimal bi-partitioning of a graph that minimizes the value of the cut. As an example of such proposals, we can cite a work that focuses on solving problems of perception of groups [75]. Instead of relying on local features and their consistencies in the image data, it tries to extract the overall impression of the image. Segmentation, seen as a graph partitioning problem, proposes a global criterion, the normalized cut, which cuts the value divided by the total weight of edges within each subgraph, for segmenting the graph.

To complete this classification, *learning based* techniques are found [76, 95]. These proposals are divided in two classes, depending on whether learning is supervised or unsupervised. The first category involves the *supervised learning based methods*. These methods select a small set of pixels in different regions to serve as prior knowledge when training a classifier. The remaining pixels are considered as the test set being partitioned into several significant regions once the classifier has been trained. The second category includes the *methods based on unsupervised learning*. Firstly, these methods estimate the optimal number of regions in the image using various indicators. Then, they apply different clustering algorithms (like k-means [52] or pose maps [86]) to partition the image into the number of regions estimated in the first step of the methods. Unsupervised learning methods are used for semantic image search, multiband automatic annotation and segmentation. In Table 2 you can find a summary of the main techniques used in object segmentation. Next to each method its main features are offered as well as some examples of implementations that can be found in the literature.

3.2. Tracking

Tracking can be defined as the problem of estimating the trajectory of an object in the image plane as it moves over a scene. In other words, tracking means assigning labels to the tracked objects in different frames of a video sequence. This poses several problems which object tracking algorithms must face: (i) loss of information caused by the projection of the real world (three dimensional) in a two-dimensional image, (ii) noise in images, (iii) complex movements of objects, (iv) non-rigid or articulated objects, (v) partial and complete occlusions of objects, (vi) changes in scene lighting, and (vii) real-time processing requirements. In a tracking scenario, an object can be defined as any item that is of interest for further analysis. Objects can be represented by forms and appearances. The performances of the most commonly applied are:

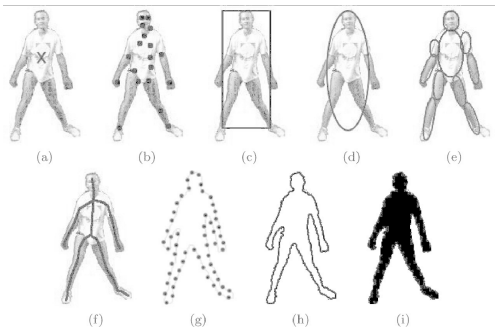


Figura 1: Representations of objects used in the tracking algorithms [90]: (a) Centroid, (b) multiple points, (c) rectangular contour, (d) elliptical contour, (e) based on multi-part forms, (f) skeleton of the object, (g) contour of the object, (h) control points on the contour of the object, (i) silhouette of the object.

- *Points*: The object is represented by a point such as the centroid (see Fig. 1a) [82], or through a set of points (see Fig. 1b) [73]. In general, this representation is suitable for tracking objects that occupy small regions in an image.
- *Primitive geometric shapes*: The shape of the object is represented by a rectangle, ellipse (see Fig. 1c and Fig. 1d [20]), etc. The movement of the object to this representation is usually modelled by translation, affine or homographic transformation. Although the primitive shapes are more appropriate to represent simple rigid objects, they are also used to represent non-rigid objects.
- *Articulated shape models*: Articulated objects are composed of body parts that are held together by connections (see Fig. 1e).
- *Skeleton models*: These models are commonly used as a representation of how to recognize objects [2], and can be used to model both articulated and rigid objects (see Fig. 1f).
- *Silhouette and contour of the object*: The representation defines the boundary contour of an object (see Fig. 1g and Fig. 1h). The region inside the boundary is known as the silhouette of the object (see Fig. 1i). The representations of the shape and contour are suitable for tracking non-rigid complex shapes [89].

We can find different proposals in the literature to represent the types of object tracking methods [90]. Among them are those based on *probability densities of the object's appearance*. In these techniques, the characteristics relating to the appearance of objects are calculated from the application of models of form over regions of the image. These models can be parametric, such as Gaussian [94], a mixture of Gaussians [67], or nonparametric [25], as Parzen windows and histograms [20]. We also have proposals based on *templates* formed by using simple geometric shapes or silhouettes [28]. These techniques provide both spatial and appearance information concerning the objects. This imposes a constraint on their time, since the templates only encode the appearance of objects generated from a single view. Therefore, they are only suitable for tracking objects whose poses do not change too much during the follow-up. On the other hand, there are techniques based on *active appearance models* [24]. These are generated simultaneously by modelling the object shape and appearance. In general, the object's shape is defined by a number of brands. In a manner similar to the contour-based representations, brands can be positioned either on the object's boundaries and within the region of the image associated with the object. For each brand a vector of appearance that may consist of a colour, texture or gradient magnitude is stored. These models require a training phase to establish both the phase and the appearance associated with the objects. Finally we find proposals based on *multiview appearance models*. These models encode different views of an object. There exists proposals for multiview appearance models for the representation of the shape and appearance of objects based on principal component analysis (PCA) and independent component analysis (ICA) [61, 6]. To summarize, Table 3 contains the main features of the previous proposals.

3.3. Interpretation of Events and Activities

The tasks associated with the interpretation of events and activities are at the highest level within the monitoring and activity detection systems. The level is considered high because it is based on tasks such as segmentation, tracking and classification, although there are proposals where these lower-level tasks are also proposed for the interpretation of events. Today, thanks to an increasing market there are affordable cameras and sensors. Systems

Cuadro 3: Proposals to represent tracking.

Proposals	Features	Publications
Probability density	Application of shape models on image regions	[94], [67], [25], [20]
Templates	Use of geometrical shapes or silhouettes Appropriate if poses do not change too much	[28]
Active appearance models	Simultaneous modelling of shape and appearance	[24]
Multiview appearance models	Generation of a subspace from the views	[61], [6]

for monitoring and activity interpretation have become popular, finding a multitude of research projects dedicated to smart surveillance, video data indexing, monitoring of the elderly and human-computer interaction through gesture recognition. Some of these projects are: ETISEO [63], AVITRACK [7], ADVISOR [1], BEWARE [5] or VSAM [18], among others.

In the literature we find many proposals where the authors use different names for the similar concepts. Thus, words like activity, behaviour and action are often used to define the same concept. Despite multiple attempts, to date, the problem of event detection is still open and challenging. This is due to several main reasons, namely, noise and the uncertainty generated by the lower levels (detection and tracking), a large variety of events, the similarity between different events, and the ambiguity to formalize an event. A list of the latest techniques to find a common ground and to create a common terminology has been provided [50]. A classification of methods for modelling events needs to be broad enough to embrace all events. In [50] the event modelling methods are classified into three main categories, uniting the multitude of proposals found in the literature. These categories are "pattern recognition methods", "state models" and "semantic models", considered the most widespread classification.

Pattern recognition methods transform the problem of representation of events into a problem of recognition and classification. The main advantage of these methods are their simplicity to be implemented and that they can be specified from the training data. In contrast, the specification does not include a semantic classifier. The most widespread methods are the "nearest neighbour", "support vector machines" and "neural networks". In comparison to other pattern recognition proposals we can highlight their simplicity and the ease with which it can be mathematically formalized and implemented. On the contrary, although it allows learning from a model data, no semantic feature is used to incorporate high level knowledge, so they are used primarily in recognition of atomic events.

The state models use semantic knowledge about events of the video (in time and space). The main improvement over pattern recognition methods is that they are able to model the state space domain, capturing the hierarchical nature and the temporal evolution of the states. Thus, these proposals allow human intuition combined with machine learning techniques. Among the state model proposals are: finite state machines, Bayesian networks, hidden Markov models, dynamic Bayesian networks and conditional random fields.

We have seen how a variety of events can be described as a sequence of states, but there are others for whom it is more appropriate to define semantic relationships between sub-events that compose them. In these cases, semantic models are suitable, as defined by the events from semantic rules, constraints and relationships, which is closely related to the way humans describe events. Thus, the problem of recognizing an event is reduced by explaining the observations using the semantic knowledge available. Such knowledge allows capturing high-level semantic relationships such as long-term time dependency, hierarchy, concurrency, and other more complex relationships between events. In contrast, the high-level nature of these models necessitates manual specification of knowledge by an expert, so the structure of the model and its parameters may not be completely defined. Within this category there are four main groups: grammar based models, Petri nets, models based on semantic constraints and logical models.

4. Multisensor Fusion

The levels previously studied are considered to be traditional in the implementation of monitoring and activity interpretation systems. These can be seen as a series of ascending levels starting from the lowest level of information acquisition. Each processing level complements the next higher level, providing all the data needed at the new level (see Fig. 2). Moreover, when developing a complete application, we must also reflect the need to develop user interfaces, as well as the control logic that orchestrates the entire monitoring process and supports data management at the different levels. However, the rise of multisensor systems, where different technologies and cameras that capture in several spectra work together with sensors of different nature (volumetric, laser, infrared), highlights the need to work together with information from all these sources. This improves the robustness of the information as it provides redundancy when obtaining data from various sources, as well as the ability to detect and correct failures in the sensors. It is therefore necessary to define and study the various multisensor fusion techniques that exist today, paying particular attention to those in the monitoring field.

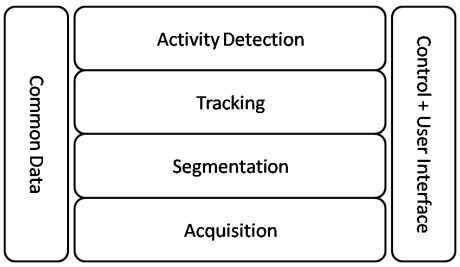


Figura 2: Traditional processing levels.

Data fusion has to be understood as the theories, techniques and tools used to combine data acquired from sensors or derived from a common representation format. According to [60], the general concept of multisensor fusion is analogous to the way humans and animals use their senses, their experience and knowledge, combining them to improve their chances of survival. Thus, multisensor fusion attempts to determine the best procedure for combining all data entries. The use of probabilistic models in fusion has the advantage of handling the uncertainty inherent in the relations between sensors and sources. For example, the Bayesian methodology, which allows fusion problems to be formulated mathematically, is capable of handling the uncertainty of the problems. One particular advantage of data fusion is that using more than one sensor it is possible to increase the quality of information obtained in several ways: increasing the spatial and temporal coverage, increasing the robustness against failures of sensors or algorithms, or better suppression of noise and more accurate estimation. In a broader sense, it is considered that data fusion improves the performance of a system in the following ways [3]:

- Representation: The information has a level of abstraction or granularity greater than the sum of the input data.
- Accuracy: Given the sensor data before the fusion S, the probability of sensory information after the fusion $p(SF)$ is greater than the prior probability of the data before the fusion $p(S)$: $(p(SF) > p(S))$
- Precision: The standard deviation of the data after the fusion is lower than that obtained directly from the sources. If the input data are noisy or erroneous, the fusion attempts to reduce or eliminate their effect.
- Completeness: After including new information to the knowledge possessed about an environment, it generates a more complete view of it.

Fusion techniques typically are designed ad-hoc, lacking a shared model that allows the realization of cooperative reasoning for event management. In [62] we find a solution to the problem, as the authors propose a data model for security systems that allow queries and introduce knowledge about security incidents, as well as about the context in which they occur. Among the problems are also the choice of an optimal multisensor fusion. We can find many works in literature which propose a series of metrics to evaluate different fusion algorithms, some well-known, such as Euclidean, Mahalanobis or correlation distance, as well as their own metrics used for comparisons in the experiments. We can find works based on genetic algorithms [4], "AND" and "OR" rules [79], or combinatorial analysis [39]. In summary, we can say that the main motivation for performing multisensor fusion is a substantial improvement in the quality of information provided by the sensors, although this improvement is dependent on the fusion technique used according to the input parameters.

5. Frameworks for Monitoring and Activity Interpretation

In the previous section the main characteristics of the different levels that make up a monitoring and activity interpretation systems have been defined. On this basis, it will be easier to understand the functioning of the different proposals of frameworks for monitoring and activity interpretation found in the literature. In this section, we introduce a series of monitoring systems, with special emphasis on surveillance systems, which are the most widespread.

5.1. Frameworks for Centralized Monitoring

The works listed below are grouped according to a centralized processing scheme. Therefore, these systems possess a single processing module that collects sensor data in a centralized manner. Fig. 3 provides a schematic representation of the systems' processing. In most theoretical approaches to monitoring frameworks are the work of [85], which proposes a programming model for the design of surveillance systems for ports and maritime

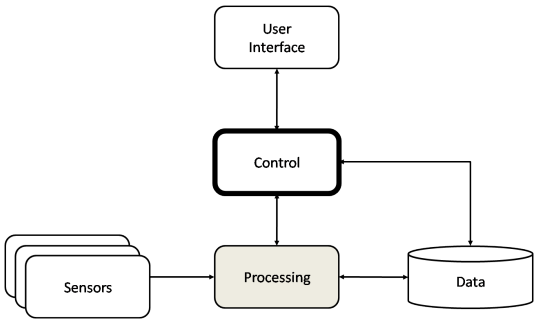


Figura 3: Centralized monitoring.

security. The model takes into account the environmental conditions, the possible combinations of sensors due to their positions and places to monitor, setting the sensor type, number and location necessary to meet system requirements while minimizing costs. This defines a decomposition strategy based on "Branch and Price" which is a decomposition of the problem in a tree by three rules for the generation of branches and two heuristics, which generate a solution to the tree.

We also have the IBM project called *Smart Surveillance System (S3)* [65], where the authors propose a middleware for use in surveillance systems that provides video-based analysis of behaviours. The system consists of two parts, the Smart Surveillance Engine (SSE) performing the video processing and analysis, and a Middleware for Large Scale Surveillance or MILS for information management. Among the features of *S3* are: (1) real-time alerts based on site, including motion detection, directional movement, abandoned objects and collecting and tampering of cameras and alerts designed by the user through an SQL-like interface, (2) search for Web-based events based on object types, sizes, speed, position, colour, length of the path and other inquiries made by the above, and (3) statistics Web-based events including times and distributions. For this, *S3* operates in the following stages: detection of moving objects, tracking of multiple 2D objects (even with occlusions), classification of objects regardless of the point of view, 3D object tracking by commercial cameras, multiscale tracking through the use of mobile pant, tilt, zoom (PTZ) cameras, object tracking across the field of vision of multiple cameras, face recognition, XML representation of objects and their attributes, indexed events in real-time and generation of Web interfaces for searching and retrieving information. This architecture has major advantages such as easy extensibility, allowing run-time addition of video sources and scalability through the use of commercial technologies (COTS).

In this line we also find the *DETEC* surveillance system [23]. The motion detection system allows the disk storage of the events (the images and associated time), associated with objects in the scene. In [38] a system is proposed for traffic control and monitoring through a single fixed camera. The system is capable of performing a thorough categorization of vehicles, improving the representation and using "linearity" (lines that make up the shape of the vehicle), so that two vehicles with the same size can be differentiated. Monitoring is done through using a Kalman filter. This system is also able to deal with shadows that can cause occlusions by using an algorithm to detect the lines dividing the lanes. Following this line we find the work of [45] which proposes a monitoring system to detect traffic accidents at intersections. To this end, a system with a single camera, where occlusions are detected by an algorithm based on Markov random fields and a space-time incident detection system is based on a hidden Markov model, combines features such as vehicle speed, direction and distance to other objects.

5.2. Frameworks for Distributed Monitoring

From now on, we call *distributed frameworks* to frameworks where processing and storage is done remotely and there is the possibility that a control unit orchestrates the processing [80]. Fig. 4 provides a simplified view of the functioning of these systems. Such architectures consist of a series of independent processing units which connect the sensors. Each of these processing units has an information storage module. Optionally, these architectures have a control module to orchestrate the distribution of information, although there are proposals where this distribution is made between the processing units themselves without orchestration. These processing units can be directly connected to users to understand the state of the unit or generate a global view of the area to monitor, obtained through the transmission of information previously discussed. In the case of the existence of a control module, it can also be used for remote maintenance tasks: to assess the status of processing modules, calibration, etc.

In [71] a visual surveillance system is proposed for tracking vehicles and pedestrians in parking lots where interactions between objects are identified. This system consists of two modules able to visually identify and track

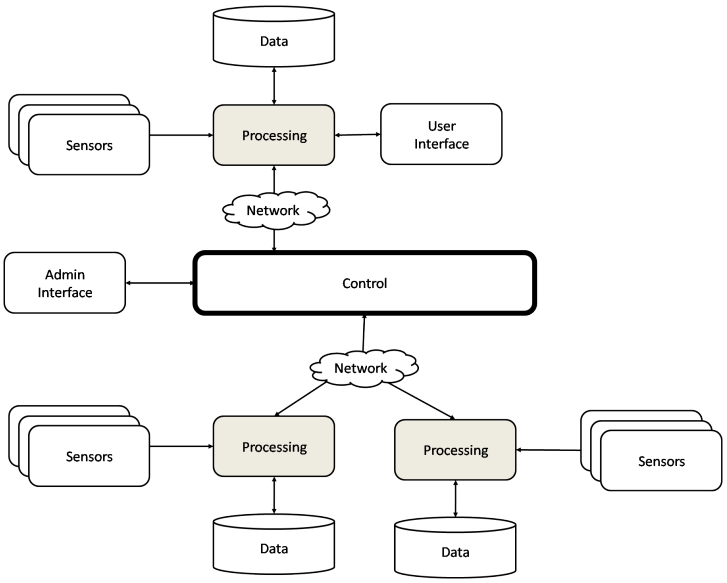


Figura 4: Distributed monitoring.

vehicles and pedestrians. Other systems such as the one described in [42] attempt to identify events of interest through the semantic interpretation of the events that take place in parking lots. This particular system consists of three parts: a tracking module, an event generator and a module capable of analysing the detected events from stochastic context-free grammars. Another surveillance system for parking can be found in [43], which seeks to identify and describe the behaviour of vehicles. For this a motion module, a model for identifying objects and a semantic system for monitoring and interpretation of the trajectories are implemented. In this work, the system builds a background model combining colour and lighting of the pixels. By calibrating the camera, it is capable of projecting a 3D model of the tracked vehicles on a plane using extended Kalman filters. After looking at the distributed monitoring systems described so far, we see that they all have two common components: a control unit to manage a variety of data processing units and a database to manage information. Although this could be considered a general line in the design, in [69] a proposal for distributed surveillance systems is described with two separate control units and nine geographically peripheral processing stations, which are installed in addition to traditional CCTV systems. This system is designed to attract the attention of the human operator if it detects certain predetermined conditions.

Yet another system is ADVISOR [1]. Its purpose is to detect events of interest in subway stations from CCTV images from the movement detected in global events for the scene. This system has been implemented using commercial hardware and open source software, all connected to a network of high bandwidth for video distribution. ADVISOR architecture consists of a number of satellite stations, located in subway stations, capable of detecting, tracking and analysing behaviour, forming a network of independent processing nodes, each with a processing unit and its own database. In [92] IVSS (English-based intelligent video surveillance system visual) is presented. A client / server architecture similar to ADVISOR is used for detection, recognition and tracking of pedestrians and vehicles. This system does not perform fusion of information at the cameras. The classification is done by support vector machines using the features extracted by a Gabor filter[1] of the blobs in the images. Another well-known distributed architecture is the VSAM project [19], where a system is organized as a network of multi-intelligent sensors. These sensors are capable of processing real-time input images for detection, monitoring and analysing events. This system also adds the ability to fuse information, generating a 3D projection of the position of objects. VSAM also provides an operator interface that displays a synthetic view of the environment where the detected objects are represented (humans and vehicles). In [51] there is another distributed monitoring architecture consisting of a series of nodes that capture through a mesh network topology, connected to the Internet which, in turn, is connected to a video processor (in charge of processing and filtering) and a monitoring station. The processing node transmits only to the monitoring station in the event that a suspicious event is detected.

[1]Gabor filters are linear filters used for edge detection. Its advantage is that is that the Gabor functions are located both in the spatial domain and in frequency, unlike the case with sinusoidal functions, which are perfectly located in the frequency domain and completely de-localized in space.

Another proposal for distributed system is found in [15], which ignores the presence of a central server, with all subsystems completely self-contained, communicating all together. Thus, the processing moves to the nodes read from the camera, avoiding the disadvantages of centralized systems where, if the central server fails, the entire system fails. There is also information fusion is performed in some nodes through a specific protocol. The nodes of this system have a camera, a processor, a frame grabber, a network interface and one database. We can see similarities in the work of [88], where several houses share the task of monitoring people. For this, establishing a priority system modelled by Markov chains used by the overload probability of a camera and an object to be rejected (e.g. the camera cannot track an object if there is an occlusion), thus establishing a dynamic load sharing of the cameras. We also find project proposals such as NICTA Open SensorWeb Architecture [16] project, which aims to integrate sensor networks with distributed computing platforms (Grids). The range of applications for the proposed architecture is wide, ranging from tsunami detection and monitoring of coral reefs to the monitoring of transport and roads or even the monitoring of air pollution. We can find a special feature of this architecture in [17], which sets out four layers, starting with the physical layer where the sensors, continuing the interaction layer with the sensors responsible for obtaining the same data, a layer of service that translates the sensor data in a message *XML*, acting with the middleware application layer, which performs information processing. In [68] a co-operative camera network (CCN) is described. An indoor surveillance system consists of a network where each node consists of a PTZ camera, a computer and a console to interact with the user. This system was developed for monitoring humans in malls to prevent theft. Another proposal [56] presents a distributed monitoring system for monitoring outdoor activities trying to learn using a probabilistic model, establishing links between the different cameras. This work allows automatic calibration of cameras enabling to relate the objects in the scene with different cameras using the number of objects that have come and gone from the scene. The inputs and outputs are modelled by a Gaussian mixture model (GMM). With this, the system can track objects even through the cameras' blind spots without calibration.

5.3. Frameworks for Hybrid Monitoring

The literature also offers proposals where part of the processing is done distributed, usually the lower-level tasks, whilst a central system is in charge of the highest-level ones. Such systems are included within the so-called *hybrid frameworks* (see Fig. 5). In [58] an overhead system is presented for the analysis of traffic where an embedded signal processing hardware (e.g. 13 processors connected to 13 cameras) and processed data are sent to a control unit that performs a post processing of the data. In this paper the authors use models based on features to track cars throughout the areas under surveillance. A work that follows the same line is found in [37] which uses a two-tiered system for the detection and monitoring of cyclists and pedestrians. The first level is a module for monitoring capable of real-time processing the image captured by the cameras, while the second level, in charge of off-line analysis, is located on a remote PC. Monitoring is done using Kalman filters. In [72] we find a CCTV surveillance system of train stations where the various modules are organized hierarchically. In this hierarchy we have a central module in charge of collecting and displaying images from all cameras. It also allows image storage and alarm generation from images sent by the remote modules. Following along the lines of these systems, we find PRISMATICA [54], a monitoring system which combines multisensor surveillance cameras with audio sensors, forming a hybrid monitoring system applied to public transport. This system consists of a number of distributed processing systems (CCTV, IP cameras, audio sensors and card readers) and a central node that collects data from legacy systems and acts as an interface with the operator. The processing is based on the Modular Integrated Pedestrian Surveillance (MIPS) architecture [53], which provides data access to database, event recording, video acquisition and display, and the processing itself.

The work of[13] shows an intrusion detection system by integrating heterogeneous information sources. The system collects information from cameras, microphones and sensors to detect situations of interest (unauthorized access) and reports these alarms in real-time via mobile devices. For this, the system has translation modules for detecting and tracking objects in images, sound detection in the case of microphones, or the activation of the sensors to generate events, and a unit processing that receives events and updates the object storage. In [81] we find a multiagent system for service-oriented monitoring aimed at improving the scalability, robustness and security. Service-oriented processing is used to allow autonomy in processing and communication system components. For scalability, we have agents responsible for adding services to the surveillance system. In the proposed architecture a layer in charge of capturing perceptual environment information (video, audio and other sensors) is provided. In this layer are remote processing nodes, which generate a series of events considered as the state change in the perception of the scene to watch. Moreover, this layer has a perceptual model of the environment. On the other hand, a conceptual layer, located in a central node, is presented, where different agents are responsible for processing the information from the previous layer, generating alarms when necessary. Therv will be a number of agents in charge of recording the events, using injected human knowledge to identify simple behaviours. Finally an agent named "Security Guard" processes the simple behaviours to identify these

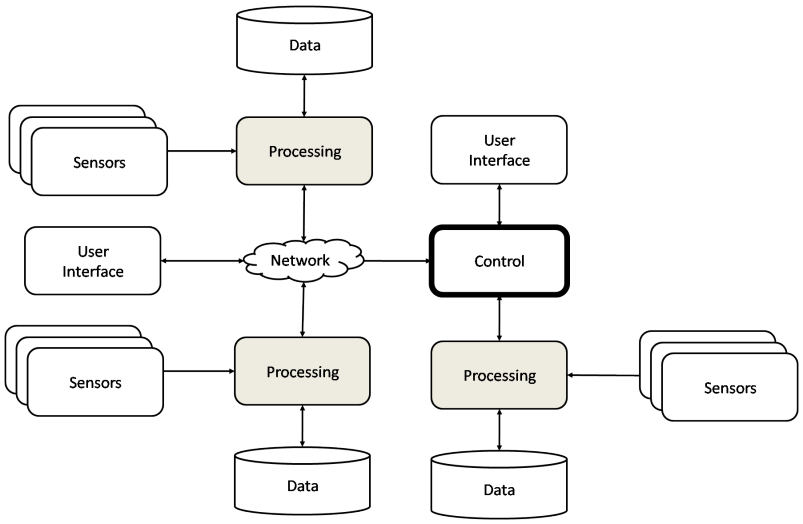


Figura 5: Hybrid monitoring.

behaviours as complex behaviours, sending them to the layer of decision-making. Another example can be found in [27] where a framework for monitoring and activity detection. The framework proposes a series of remote nodes that perform low-level processing (segmentation and tracking) and a central node that gathers information from the remote ones, and performs information fusion and activity detection. This framework has been used for a range of tasks such as fall detection [10] or recognizing and regulating emotions [11, 12]. [47] describes a multisensor architecture for event detection in video sequences where the focus changes from one sensor to another for monitoring as optimally as possible. The events are detected by finite state machines. The architecture is based on a client web server where a database stores the events. The proposal includes video processing, event detection and extraction of metadata modules for different types of sensors (video, audio and still images). The metadata are transformed into events that are presented to the user. The detection is done by subtraction of images and, in order to detect events, the system uses training sets to determine the spatial and temporal characteristics of the events. In [87] we find a system of classification and tracking of players with eight cameras, each with its own processor for segmentation and classification and monitoring. A central processor collects this data and fuses them using the nearest neighbour method by Mahalanobis distance [55]. For the monitoring purpose Kalman filters are used.

As a summary of these hybrid systems, a set of common components in their design has been found. First, they have a number of separate processing modules where the sensors are connected, as was the case with distributed systems. The difference is that these modules do not perform distributed processing at all levels. For that a control unit connected directly to a processing module level, a storage module and other user interfaces are used. Thus, the remote modules perform tasks such as segmentation and tracking of objects and the processing module handles high-level tasks of information fusion and remote sensing of events to generate a global surveillance system.

6. Conclusions

This paper has reviewed various proposals for monitoring and activity interpretation frameworks that we have found in the literature. This survey was divided into four distinct parts. First, we took into account the scope of monitoring and interpretation systems, distinguishing between research-oriented proposals and those aimed at more commercial purposes. After that, in order to better understand the characteristics of current systems, we reviewed different generations of monitoring systems, from the first closed circuit television, dating to the 60s, to third-generation systems where, in addition to providing the processing capabilities of information of second generation systems, it has been proposed to perform distributed processing in order to monitor large areas with many heterogeneous sensors. Once the characteristics of monitoring systems were identified, we also wanted to pay attention to the algorithms that operate on each of the classic levels of abstraction depending on the level of information, these being detection, object tracking and activity interpretation. We also defined the main concepts associated with each of the levels for a better understanding of their functionality, again through offering ratings for the algorithms found in each of the levels according to their algorithmic basis.

It seems logical that if we are to create a multisensor system of monitoring and interpreting, we also need to

merge information from various sources. Knowing the main features of the systems for monitoring and activity interpretation, from detection and object tracking to the detection of activities, while incorporating information fusion, we have paid special attention to the frameworks for monitoring and activity interpretation that can be found in the literature. Articles in this section have been analysed taking into account that the levels proposed by these architectures serve as reference for the processing levels of the architecture to be proposed later. Unfortunately, we found that most articles, even though in many of them the word "framework" is used, usually propose specific solutions to a given problem and not the more generic solutions you would expect from a proper architecture. Still, we have classified the different techniques based on several criteria such as type of organization of the processing (centralized, distributed and hybrid or hierarchical).

Finally, after the study of the different techniques offered in this paper, we highlight some of the benefits that they provide to obtain an architecture for monitoring and interpretation. Modularity should be noted first, allowing the system to be scalable depending on the needs of it. So, the inclusion of new modules has to be allowed without the implementation of major changes in the system. On the other hand, it is important that the architectures operate on several levels where fusion levels intertwine with traditional levels of processing to maximize the benefits of information sources. The importance of modelling has also to be highlighted as a fundamental part not found in most architectures. The model can refer to both the environment and the sensors. For large systems, where multiple sensors cooperate to perform the monitoring tasks, it is necessary to set some parameters to identify objects throughout the scene, regardless of the sensors that detect them. This is also necessary to establish a model of execution, first to download all the processing nodes, and secondly to obtain an overview of the scenario to be monitored. After studying the different models, it has been seen that the execution model that best fits these requirements is the hybrid or hierarchical one.

To summarize we can draw some design trends of the current monitoring architectures. Firstly, the use of a large number of sensors connected to processing nodes to cover big areas to monitor in real-time and secondly, to attract the attention of the human operator when an event of interest is detected. These nodes are able to process information independently, allowing for greater scalability and robustness.

Acknowledgment

This work was partially supported by Spanish Ministerio de Economía y Competitividad / FEDER under DPI2016-80894-R grant.

Referencias

[1] ADVISOR. http://www-sop.inria.fr/orion/advisor/, 2003. Accessed 26 May 2016.

[2] A. Ali and J.K. Aggarwal. Segmentation and recognition of continuous human activity. In *Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on*, pages 28 –35, 2001.

[3] David Bellot, Anne Boyer, and Francois Charpillet. A new definition of qualified gain in a data fusion process: Application to telemedicine. In *Proceedings Fifth International Conference on Information Fusion.*, Annapolis, Maryland, July 2002.

[4] Wael Ben Soltana, Mohsen Ardabilian, Liming Chen, and Chokri Ben Amar. Optimal fusion scheme selection framework based on genetic algorithms for multimodal face recognition. In *Compression et Representation des signaux audiovisuals (CORESA),*, October 2010.

[5] BEWARE. http://www.eecs.qmul.ac.uk/ sgg/beware/, 2011. Accessed 26 May 2016.

[6] Michael J. Black and Allan D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26:63–84, January 1998.

[7] Mark Borg, David Thirde, James Ferryman, Florent Fusier, François Bremond, and Monique Thonnat. An integrated vision system for aircraft activity monitoring. In *In Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS05*, 2005.

[8] Hilary Buxton. Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, 21(1):125 – 136, 2003.

[9] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:679–698, November 1986.

[10] José Carlos Castillo, Davide Carneiro, Juan Serrano-Cuerda, Paulo Novais, Antonio Fernández-Caballero, and José Neves. A multi-modal approach for activity classification and fall detection. *International Journal of Systems Science*, 45(4):810–824, 2014.

[11] José Carlos Castillo, Antonio Fernández-Caballero, Álvaro Castro-González, Miguel A Salichs, and María T López. A framework for recognizing and regulating emotions in the elderly. In *International Workshop on Ambient Assisted Living*, pages 320–327. Springer International Publishing, 2014.

[12] José Carlos Castillo, Angel Rivas-Casado, Antonio Fernández-Caballero, María T López, and Rafael Martínez-Tomás. A multisensory monitoring and interpretation framework based on the model–view–controller paradigm. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 441–450. Springer Berlin Heidelberg, 2011.

[13] J.L. Castro, M. Delgado, J. Medina, and M.D. Ruiz-Lozano. Intelligent surveillance system with integration of heterogeneous information for intrusion detection. *Expert Systems with Applications*, 38(9):11182–11192, September 2011.

[14] Zhengying Chen, Yonghong Tian, Wei Zeng, and Tiejun Huang. Detecting abnormal behaviors in surveillance videos based on fuzzy clustering and multiple auto-encoders. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[15] M. Christensen and R. Alblas. V2-design issues in distributed video surveillance systems. Technical report, Department of Computer Science, Aalborg University, Aalborg, Denmark, 2000.

[16] X. Chu, T. Kobialka, B. Durnota, and R. Buyya. Open sensor web architecture: Core services. In *Intelligent Sensing and Information Processing, 2006. ICISIP 2006. Fourth International Conference on*, pages 98–103, Oct 2006.

[17] Xingchen Chu and Rajkumar Buyya. Service Oriented Sensor Web. *Sensor Networks and Configuration*, pages 51–74, 2007.

[18] Robert Collins, Alan Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, and Osamu Hasegawa. A system for video surveillance and monitoring. Technical Report CMU-RI-TR-00-12, Robotics Institute, Pittsburgh, PA, May 2000.

[19] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456 –1477, October 2001.

[20] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564 – 577, May 2003.

[21] Timothee Cour, Florence Benezit, and Jianbo Shi. Spectral segmentation with multiscale graph decomposition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1124–1131. IEEE, 2005.

[22] R. Cucchiara, C. Grana, A. Prati, G. Tardini, and R. Vezzani. Using computer vision techniques for dangerous situation detection in domotic applications. In *Intelligent Distributed Surveilliance Systems, IEE*, pages 1 – 5, Feb 2004.

[23] Detec. http://www.detec.no, 2011. Accessed 26 May 2016.

[24] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, FG '98, pages 300–, Washington, DC, USA, 1998. IEEE Computer Society.

[25] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151 – 1163, July 2002.

[26] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[27] Antonio Fernández-Caballero, José Carlos Castillo, María T López, Juan Serrano-Cuerda, and Marina V Sokolova. Int 3-horus framework for multispectrum activity interpretation in intelligent environments. *Expert Systems with Applications*, 40(17):6715–6727, 2013.

[28] Paul Fieguth and Demetri Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 21–, Washington, DC, USA, 1997. IEEE Computer Society.

[29] Loren Fiore, Duc Fehr, Robot Bodor, Andrew Drenner, Guruprasad Somasundaram, and Nikolaos Papanikolopoulos. Multi-camera human activity monitoring. *Journal of Intelligent & Robotic Systems*, 52:5–43, 2008. 10.1007/s10846-007-9201-6.

[30] Michael Fleischman, Phillip Decamp, and Deb Roy. Mining temporal patterns of movement for video content classification. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 183–192, New York, NY, USA, 2006. ACM.

[31] BE Furby and BD Roney. Autonomous surveillance in the visual spectral region. In *In Materials Research Labs. Extracts from Symp.: Countersurveillance 1983 p 48-68 (SEE N84-34779 24-43)*, volume 1, pages 48–68, 1984.

[32] José M Gascueña and Antonio Fernández-Caballero. On the use of agent technology in intelligent, multisensory and distributed surveillance. *The Knowledge Engineering Review*, 26(02):191–208, 2011.

[33] Zeno Geradts and Jurrien Bijhold. Forensic video investigation. In Gian Luca Foresti, Petri Mahonen, and Carlo S. Regazzoni, editors, *Multimedia video based surveillance systems*, chapter 1, pages 3–12. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[34] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing 3rd Ed.* Prentice-Hall, Englewood Cliffs, NJ., 2007.

[35] Juan Gómez-Romero, Miguel A. Serrano, Jesús García, José M. Molina, and Galina Rogova. Context-based multi-level information fusion for harbor surveillance. *Information Fusion*, 21:173 – 186, 2015.

[36] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):809 –830, August 2000.

[37] J. Heikkila and O. Silven. A real-time system for monitoring of cyclists and pedestrians. In *Visual Surveillance, 1999. Second IEEE Workshop on, (VS'99)*, pages 74 –81, July 1999.

[38] Jun-Wei Hsieh, Shih-Hao Yu, Yung-Sheng Chen, and Wen-Fong Hu. Automatic traffic surveillance system for vehicle tracking and classification. *IEEE Transactions on Intelligent Transportation Systems*, 7(2):175 – 187, 2006.

[39] D.F. Hsu, D.M. Lyons, and J. Ai. Selecting and evaluating combinatorial fusion criteria to improve multitarget tracking. In *9th International Conference on Information Fusion*, pages 1 –7, july 2006.

[40] Ipsotek. http://www.ipsotek.com, 2011. Accessed 26 May 2016.

[41] Kashif Iqbal, Michael Odetayo, Anne James, Rahat Iqbal, Neeraj Kumar, and Shovan Barma. An efficient image retrieval scheme for colour enhancement of embedded and distributed surveillance images. *Neurocomputing*, 174, Part A:413 – 430, 2016.

[42] Y. Ivanov, C. Stauffer, A. Bobick, and W.E.L. Grimson. Video surveillance of interactions. In *Second IEEE Workshop on Visual Surveillance (VS'99)*, pages 82 –89, July 1999.

[43] L. Jian-Guang, L. Qi-Feing, T. Tie-Niu, and H. Wei-Ming. 3d model based visual traffic surveillance, acta automatica sinica (chinese. *Journal of Automation, Chinese Academy of Sciences*, 29(3):434–449, 2003.

[44] Cláudio Rosito Jung and Jacob Scharcanski. Robust watershed segmentation using wavelets. *Image and Vision Computing*, 23(7):661 – 669, 2005.

[45] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi. Traffic monitoring and accident detection at intersections. *IEEE Transactions on Intelligent Transportation Systems*, 1(2):108 –118, jun 2000.

[46] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 1(4):321–331, 1988.

[47] Declan Kieran and WeiQi Yan. A framework for an event driven video surveillance system. *Journal of multimedia*, 6(1):3–13, Feb 2011.

[48] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *Third IEEE International Workshop on Visual Surveillance*, pages 3 –10, 2000.

[49] Manish Kushwaha, Songhwai Oh, Isaac Amundson, Xenofon Koutsoukos, and Akos Ledeczi. Target tracking in urban environments using audio-video signal processing in heterogeneous wireless sensor networks. In *Proc. of the 42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, October 2008.

[50] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(5):489 – 504, Sept 2009.

[51] Francesco Licandro and Giovanni Schembra. Wireless mesh networks to support video surveillance: architecture, protocol, and implementation issues. *EURASIP J. Wirel. Commun. Netw.*, 2007:34–34, January 2007.

[52] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[53] B. Lo, S.A. Velastin, M.A. Vicencio-Silva, and Jie Sun. An intelligent distributed surveillance system for public transport. *IEE Seminar Digests*, 2003(10062):10–10, 2003.

[54] Benny Ping Lai Lo, Jie Sun, and Sergio A. Velastin. Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems. *Acta Automatica Sinica*, 29(3):393–407, May 2003.

[55] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936.

[56] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–205 – II–210 Vol.2, July 2004.

[57] R. Martínez-Tomás, M. Rincón, M. Bachiller, and J. Mira. On the correspondence between objects and events for the diagnosis of situations in visual surveillance tasks. *Pattern Recogn. Lett.*, 29:1117–1135, June 2008.

[58] Philip McLauchlan, David Beymer, Benn Coifman, and Jitendra Mali. A real-time computer vision system for measuring traffic parameters. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, Washington, DC, USA, 1997. IEEE Computer Society.

[59] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *Computer VisionECCV 2004*, 1:69–82, 2004.

[60] H. B. Mitchell. *Multi-Sensor Data Fusion. An Introduction.* Springer-Verlag, 2007.

[61] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696 –710, jul 1997.

[62] Benjamin Morin, Ludovic Mé, Hervé Debar, and Mireille Ducassé. A logic-based model to support alert correlation in intrusion detection. *Information Fusion*, 10(4):285 – 299, 2009. Special Issue on Information Fusion in Computer Security.

[63] M. Thonnat Nghiem A.-T., F. Bremond and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *Proceedings of AVSS 2007*, 2007.

[64] ObjectVideo. http://www.objectvideo.com, 2010. Accessed 26 May 2016.

[65] Vio Onut, Don Aldridge, Marcellus Mindel, and Stephen Perelgut. Smart surveillance system applications. In *Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research*, CASCON '10, pages 430–432, New York, NY, USA, 2010. ACM.

[66] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–66, March 1979. minimize inter class variance.

[67] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(3):266 –280, March 2000.

[68] I. Paulidis and V. Morellas. Two examples of indoor and outdoor surveillance systems. In C. S. Regazzoni, G. Fabri, and G. Vernazza, editors, *Advanced Video-Based Surveillance Systems*, pages 39–51. Kluwer Academic, Boston, MA, USA, 1998.

[69] M. Pellegrini and P. Tonani. Security in ports: the user requirements for surveillance system. In C. S. Regazzoni, G. Fabri, and G. Vernazza, editors, *Advanced Video-Based Surveillance Systems*, pages 18–26. Kluwer Academic, Boston, MA, USA, 1998.

[70] Ramani Pichumani. Snakes: An active model, jul 1997.

[71] P. Remagnino, A. Baumberg, T. Grove, D. Hogg, T. Tan, A. Worrall, and K. Baker. An intergrated traffic and pedestrian model-bassed vision system. In *Proceedings of the BMVC '97*, pages 380 – 389, Israel, 1997.

[72] N. Ronetti and C. Dambra. Railway station surveillance: the italian case. In Gian Luca Foresti, Petri Mahonen, and Carlo S. Regazzoni, editors, *Multimedia video based surveillance systems*, chapter 1, pages 13–20. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[73] D. Serby, E.K. Meier, and L. Van Gool. Probabilistic object tracking using multiple features. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 184 – 187 Vol.2, 2004.

[74] Ljiljana Seric, Darko Stipanicev, and Maja Stula. Observer network and forest fire detection. *Information Fusion*, 12(3):160 – 175, 2011. Special Issue on Information Fusion in Future Generation Communication Environments.

[75] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 1997.

[76] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[77] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747 –757, August 2000.

[78] Sun-Gu Sun and HyunWook Park. Segmentation of forward-looking infrared image using fuzzy thresholding and edge detection. *Optical Engineering*, 40:2638–2645, 2001.

[79] Q. Tao, R.T.A. van Rootseler, R.N.J. Veldhuis, S. Gehlen, and F. Weber. Optimal decision fusion and its application on 3d face recognition. In A. Bromme, C. Busch, and D. Huhnlein, editors, *Proceedings of the Special Interest Group on Biometrics and Electronic Signatures*, GI-Edition, pages 15–24, Germany, July 2007. Gesellschaft fur Informatik e.V.

[80] M. Valera and S.A. Velastin. Intelligent distributed surveillance systems: a review. *Vision, Image and Signal Processing, IEE Proceedings* -, 152(2):192–204, April 2005.

[81] David Vallejo, Javier Albusac, Carlos Gonzalez-Morcillo, and Luis Jimenez. A service-oriented multiagent architecture for cognitive surveillance. In *Proceedings of the 12th international workshop on Cooperative Information Agents XII*, CIA '08, pages 101–115, Berlin, Heidelberg, 2008. Springer-Verlag.

[82] C.J. Veenman, M.J.T. Reinders, and E. Backer. Resolving motion correspondence for densely moving points. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(1):54 –72, January 2001.

[83] M.E. Weber and M.L. Stone. Low altitude wind shear detection using airport surveillance radars. In *Record of the 1994 IEEE National Radar Conference, 1994*, pages 52–57, March 1994.

[84] I.H. White, M.J. Crisp, and R.V. Penty. A photonics based intelligent airport surveillance and tracking system. *Advanced Information Networking and Applications, International Conference on*, 0:11–16, 2010.

[85] W.E. Wilhelm and E.I. Gokce. Branch-and-price decomposition to design a surveillance system for port and waterway security. *IEEE Transactions onAutomation Science and Engineering*, 7(2):316 –325, 2010.

[86] Ziyan Wu, Yang Li, and Richard J Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(5):1095–1108, 2015.

[87] M. Xu, J. Orwell, L. Lowey, and D. Thirde. Architecture and algorithms for tracking football players with multiple cameras. *Vision, Image and Signal Processing, IEE Proceedings* -, 152(2):232 – 241, April 2005.

[88] Yi Yao, Chung-Hao Chen, Andreas Koschan, and Mongi Abidi. Adaptive online camera coordination for multi-camera multi-target surveillance. *Computer Vision and Image Understanding*, 114(4):463 – 474, 2010. Special issue on Image and Video Retrieval Evaluation.

[89] A. Yilmaz, Xin Li, and M. Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1531 –1536, 2004.

[90] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38, December 2006.

[91] Zhiwen Yu, Hau-San Wong, and Guihua Wen. A modified support vector machine and its application to image segmentation. *Image and Vision Computing*, 29(1):29 – 40, 2011.

[92] Xiaojing Yuan, Zehang Sun, Y. Varol, and G. Bebis. A distributed visual surveillance system. In *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, pages 199 – 204, July 2003.

[93] E. Zervas, A. Mpimpoudis, C. Anagnostopoulos, O. Sekkas, and S. Hadjiefthymiades. Multisensor data fusion for fire detection. *Information Fusion*, 12(3):150 – 159, 2011. Special Issue on Information Fusion in Future Generation Communication Environments.

[94] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multi-band image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:884–900, 1996.

[95] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs.* Carnegie Mellon University, language technologies institute, school of computer science, 2005.

[96] Djemel Ziou and Salvatore Tabbone. Adaptative elimination of false edges for first order detectors. In *International Conference on Image Analysis and Processing*, pages 89–94. Springer, 1995.