Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Christian Micheloni · Marco Lestuzzi · Gian Luca Foresti

Adaptive Video Communication for an Intelligent Distributed System

Tuning sensors parameters for surveillance purposes

Abstract Surveillance systems includes a large set of techniques for both low level and high level tasks. In particular, in the last decade the research community has witnessed to a high proliferation of techniques that span from object detection and tracking to object recognition and event understanding. Although some techniques have been proved to be very effective, those tasks cannot be considered solved. Even less, we can consider concluded the research in the field of the analysis of the activities (event analysis). It is this topic together with the problem of the information sharing among different sensors that represent the core of this work.

Here, a system architecture for a video surveillance system with distributed intelligence over multiple processing units and with distributed communication over multiple heterogeneous channels (wireless, satellite, local IP networks, etc.) is proposed. A new real-time technique for changing the video transmission parameters (e.g., frame rate, spatial/color resolution, etc.) according to the available bandwidth (which depends on the number of the detected alarm situations, on the required video quality, etc.) will be presented.

Keywords Surveillance · Multi sensors · Transmission

1 Introduction

Remote surveillance of unattended environments (e.g., metro line platforms, railway stations, waiting rooms, airport taxiways, nuclear plants, etc.) is a typical problem implying recognition and communication aspects [8]. Recognition is involved to classify detected objects (e.g., pedestrians, groups of people, motorcycles, cars, vans, lorries, buses, etc.) moving in the observed scene [9] for understanding their behaviours [3, 17, 25] in order to detect anomalous events [7, 29]. Communication is necessary to transmit to a remote operator useful information for monitoring the observed area and, if necessary, to take the appropriate decisions [9]. Surveillance systems provided various degrees of assistance to operators and evolved in an incremental way according to the progress in surveillance technologies [20]. Several kinds of sensors are today available for remote surveillance: they range from tactile or pressure sensors (e.g., perimtrical surveillance) to chemical sensors (e.g., industrial plant surveillance) and from audio to visual sensors [3]. For the monitoring of wide outdoor areas, visual sensors are more useful as they provide more high-level information; for this reason, the focus of this paper is on those applications where visual information plays the paramount role.

Video data are acquired by distributed sources and then are usually transmitted to some remote control centre. As the number of cameras increases, remote monitoring by human operators is rather boring, tedious and error-prone. Hence, the automatic processing of the video data can help the operator in reducing the number of interesting events to be analysed. When compared to audio or text signals, videos are characterised by a higher amount of information obliges wide areas surveillance system to consider the communication layer as a really important feature. Here, the up-link (from the sensors to control centre) bandwidth plays the key role as it represents the bottleneck for the information sharing (i.e. video upload) between peripheral nodes and central operative units.

Today, the operators of visual surveillance systems are required to inspect images and video at high quality to detect and understand important information (e.g., the face of a person, the license plate of a car, etc.). The same requirement arises for automatic processing systems. The information rate required by a visual surveillance system can be fixed by considering the acquisition rate (e.g. 25 frame per second), the image size (e.g. square 256x256 or 512x512 images), the image information (full colour / black and white) and the sample quantization (e.g. 8bit/pixel/colour). In order to transmit high quality video data from large sets of visual sensors, high bandwidths are necessary. However, in several

C. Micheloni · M. Lestuzzi · G.L. Foresti
University of Udine
Tel.: +39 0432 558420
Fax: +39 0432 558499
E-mail: michelon@dimi.uniud.it

real applications a wide bandwidth is not always available, due to too high costs (e.g., non military domains) or to infrastructure limits.

Several standards have been defined in the last years for the coding of visual information. The MPEG (Moving Picture Expert Group) standards address the compression of video signals. MPEG-1 [11] operates at bit rates of about 1.5 Mbit/s and targets data transmission over communication channel as integrated-services digital networks (ISDN) or local area networks (LAN). MPEG-2 [2] operates at bit rates around 10 Mbit/s and is designed for the compression of higher resolution video signals. With the H.261 [13] recommendation data transmission at bit rates down to 64Kbit/s became possible, while the recommendation ISO/MPEG-4 [12] allows visual communication applications at bit rates low as 9.6 Kbit/s.

Several works have been done in the field of coding image sequences at a very low bit rate. A review of existing approaches to very low bit rate image and video coding can be found in [27], where these approaches have been classified into four classes being waveform, fractal coding techniques, model-based and object based. Object-based approaches are particularly suited in video-based surveillance applications that require the transmission of visual information about recognized objects carrying out dangerous situations, (e.g., a vehicle stationing on a railway crossing, an obstacle occluding an emergency lane, a person accomplishing vandalism acts, etc). In [10], Hata et al. propose a solution for an object-aware video transcoding applied to visual surveillance. The idea is to adapt the compression with respect to the objects moving in the scene. In particular, pixels belonging to the background can be more compressed than the pixels of foreground objects. The drawback of such a solution relies on the application to a single sensor that cannot be of the active type (it must be static).

The main objective of this paper is to propose a system architecture for a video surveillance system with distributed intelligence over multiple processing units and with an innovative distributed communication over multiple heterogeneous channels (wireless, satellite, local IP networks, etc.). A new real-time technique for changing the video transmission parameters (e.g., frame rate, spatial/color resolution, etc.) according to the bandwidth available (which depends on the number of detected alarm situations, on the required video quality, etc.) will be presented. In particular, when a contention system, like a distributed surveillance system certainly is, must be defined, the data that have to be sent and the way these are delivered are aspects of paramount importance. Key decisions have to be taken with respect to the principal aspects as communication protocols, compression algorithms and conflicts resolution. For what concerns the advances in communication protocols for multimedia data the available protocols, spanning from Transport Control Protocol (TCP) to TP/RTCP widely used in the last years, guarantee a real time transmission support for different types of networks like Local Area Network (LAN) or WI-FI. Also in the field of compression techniques we have witnessed to a really fast evolution of the algorithms allowing us to send really good quality multimedia data thorough low bit-rate channels. It is in the field of conflict resolution, that a bigger effort can be made. In the specific case of a video surveillance network, we have that different sensors can simultaneously send data through satellite or WI-FI channels. In this case the common techniques of controlling the data payload are not sufficient to guarantee the quality of service required by higher level decision modules installed on remote (control centers) nodes. Indeed, giving equal bandwidth to each sensor will result in a large degradation of the image quality and hence in a lower effectiveness of the decision making process who has to cope with really altered data.

Instead, considering the importance of the data that each sensor is sending into the bandwidth assignment process will result in a more efficient partition of the available bit rate. Hence, the key points of the proposed solution rely on how the data importance is assessed, on how the available bit rate can be estimated and on how these information can be exploited for the optimization of the bandwidth shares.

The remainder of the paper is organised as follows. In Section 2 the logical hierarchic architecture will be presented with particular emphasis on the data flow that from local to global gets a transformation from raw to semantic meaning. In Section 3 the way the data importance is assessed will be described with respect to the event analysis techniques that can be exploited to focus the attention of few relevant actions. In Section 4 the method developed to estimate the traffic and therefore the congestion of the network will be described together with the methods for a discriminating assignment of the sensors bit rate. Finally, in Section 5 a deep validation of the proposed techniques will be offered by showing the obtained performance on networks like Satellite and WI-FI.

2 Distributed Architecture

As previously explained, a modern video surveillance system is composed by a large amount of sensors. Hence, defining their displacement and their duty is a task of paramount importance in deciding the architecture of the surveillance system. In [18] Micheloni *et al.* proposed a hierarchic architecture where static and active sensors cooperate in a distributed manner. Such a scheme have been inherited by the proposed solution where some of its concepts like communication as consequence of semantic information have been deeply investigated and optimised.

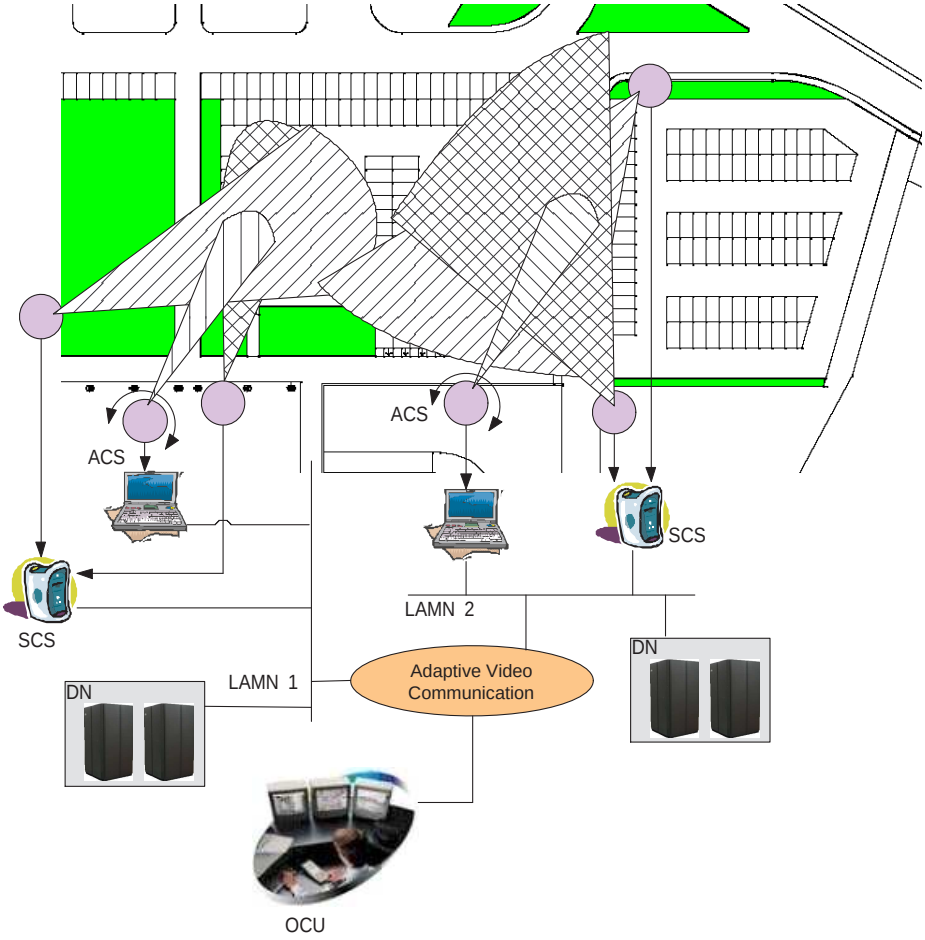

Fig. 1 System architecture. The entire sensor network is hierarchically organised in Local Area Monitoring Networks (LAMN) each responsible of the surveillance of a restricted zone. Data acquired and computed by each node inside each LAMN can be sent either within the LAN or directed to nodes in other LANs. The Operative Control Unit (OCU) is responsible of the surveillance of all the zones covered by the LAMN connected to it. For these reasons it has the necessity to receive as much information is possible from its underlying networks.

Local Area Monitoring Networks (LAMNs) are organised in a hierarchic way to provide the monitoring of local areas (see Fig. 1). Static Camera Systems (SCSs) are defined to detect and track all the objects moving within the assigned area while for gazing targets with higher resolution Active Camera Systems (ACSs) have been exploited. The sensor placement is done by balancing two opposite needs: on one hand the possibility to cover the entire area minimizing the number of sensors while on the other the need for redundancy to counteract targets occlusions by providing overlapping fields of view.

In addition to this definition, that is mostly physical, a logical definition has been provided to describe the information flow through the processing levels that compose the proposed solution (see Fig. 2). A First Level Node (FLN), either SCS or ACS, acquire raw data and perform a first step of processing by applying well know techniques as object detection, tracking, recognition etc. Such information regarding small zones (section of a parking lot, part of a station hall, train platform, etc.) area are sent within the LAMN of pertinence to a first Decision Node (DN). At this level, the information processed by FLNs within different zones are fused together to achieve a first level of decision making. In particular, more complex events can be detected and if suspicious behaviors arise inside the local area (parking lot, station hall, train departures area, etc) appropriate alarms can be generated.

As result, with respect to FLNs at the DN level more relevant features and therefore semantically meaningful information can be extracted. At this level a first skim of the entire data flow is performed in order to send only relevant information to the Operative Control Unit (OCU) which represents the highest level node (root) of the hierarchy. At this stage, all the information gathered by the entire network of sensors and the derived features are analysed in order to give a definitive response to the

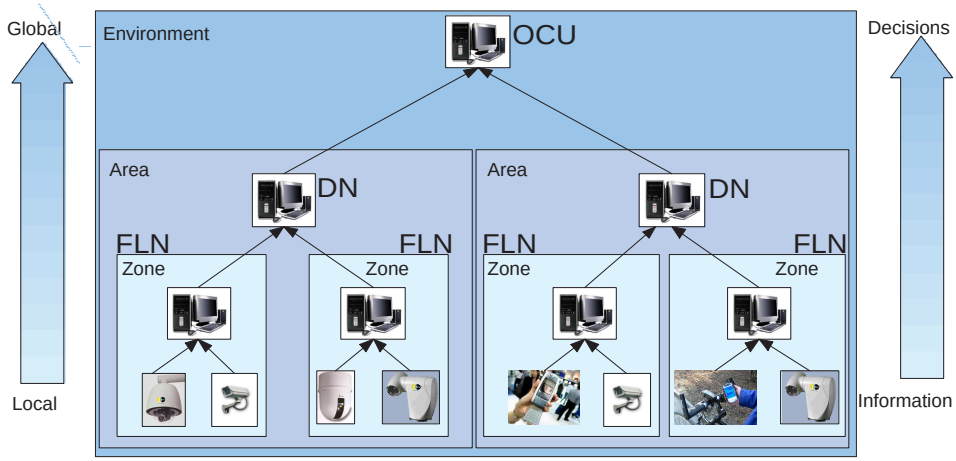

Fig. 2 Information flow scheme. The type of data flowing from the bottom to the top of the network architecture goes through several transformations that lead to a more semantic and descriptive information as it goes up to the root of the network. This is due to the fact that when each node receives the data it also processes it to extract more useful information for the decision making nodes.

actions occurring within the monitored environment. It is indeed at this node that the whole picture is available therefore allowing a better analysis of the complex and composite events for an effective decision and threat response.

To achieve the global awareness of the proposed architecture information has to flow from the low levels to upper nodes. If all acquired data was to be sent through the network, bandwidth requirements would quickly increase. As we often have to adopt band-limited channels we need to appoint a communication system for the optimization of such resources. At time intervals the upper nodes probe the network to estimate its traffic and on the basis of the importance of the sensors' content they allocate a quota of bandwidth to each underlying node. These nodes, on the basis of what is happening in their area of interest, select the best data representation to satisfy the given constraints (i.e. bandwidth quota).

3 Security Tasks

When a surveillance system has to monitor a single wide area or multiple disjoint areas, the way information between area controllers is shared is of fundamental importance. This is even more true if, as in the adopted hierarchic network, the output of each level depends on the results provided through data communication by underlying levels. In addition, as at each node we have adopted algorithms for the motion detection, object recognition and event analysis we can strengthen the reliability exploiting the outputs to better inform the upper nodes. In particular, it is possible to develop a system where nodes can cooperate to better access the communication resources with the aim of increasing the overall system robustness.

Let us now think to the case of a normal event that is happening within the field of view of a node (i.e. person is getting a car). The answer of the system based merely on such an information could be only one: no threats occurring. But, if we can put together the information extracted by another node concerning the same person who parked a car near a sensitive target, then the system response would be certainly different. For this reason, a robust monitoring of the activities taking place within the environment requires a process of information sharing within and between the single LAMN. As result of the monitoring process, each node is able to determine the level of risk about the activities happening inside its area of interest.

As consequence, it can decide to improve the analysis performance by tasking the underlying nodes, if any exists, by requiring more detailed information or to inform upper nodes about what is going on the stage. All these decisions have to be taken as consequence of the activities each surveillance system counts on. These activities will be deeply described in the following sections by explaining the solutions adopted to develop the proposed system. In particular we will give more information on three principal tasks:

- Object Detection
- Object Classification and Recognition
- Event Analysis

3.1 Object Detection

With regards to the detection of moving objects, we have based our solution on a change detection method. A variety of change detection methods have been developed either for fixed or active cameras. An exhaustive review can be found in [7]. The simple difference (SD) method

$$y_k = f\left(\sum_{j=1}^{M} w_{kj}x_j + \sum_{j=1}^{M}\sum_{i=1}^{M} w_{kji}x_j x_i + \sum_{j=1}^{M}\sum_{i=1}^{M}\sum_{l=1}^{M} w_{kjil}x_j x_i x_l\right) \quad (1)$$

computes for each time instant t the absolute difference between the pixel intensities of the input and a reference image, then it applies a threshold Th to obtain a binary image $B(x, y)$. Threshold selection is a critical task, and those proposed by Kapur [15], Otsu [19], Ridler [22], Rosin [23] have been recently outperformed both in performance and in reliability by the thresholding technique proposed by Snidaro and Foresti [24]. In the case of static cameras,to minimise errors in the change-detection process due to noise, illumination and/or environmental changes, advanced visual-based surveillance systems apply background updating procedures as part of frame-to-background image differencing techniques. Here, a multi-color model proposed by Bhanakar and Luo [1], where multiple colour clusters are used to represent the background at each pixel location, is exploited. Instead, when a camera is moving more complex operations as registration techniques have to be considered as a pre-processing step before change detection. A survey of such methods [28] explains how particular transforms (translational, affine, perspective, etc.) can be computed to solve the problem. In addition, after registering the current image and having thresholded the differences some problems have to be considered. Indeed, the frame by frame technique results in a coarse identification of the moving objects especially if their speed is slow, as shown by Collins *et al.* [4]. What happens is that holes appear inside the blobs and the detected moving pixels represent just a small part of the entire object. Such pixels are then used as seeds in a region growing process [5] which determines moving regions and therefore the blobs corresponding to the detected moving pixels. For both static and moving cameras, the resulting blobs are given to the recognition module for the object classification and thereafter for its identification.

3.2 Object Classification and Recognition

The object classification module tries to assign each detected object to a predefined set of categories (e.g. cars, pedestrians and cycles in outdoor environments). In the proposed system, object classification is performed by means of an adaptive high-order neural tree (AHNT) classifier [6]. The AHNT is a hierarchical multi-level neural network, in which the nodes are organized into a tree topology. The AHNT is built by successively dividing the training set into subsets (local training set) and assigning each subset to a different child node. Nodes are high-order perceptrons [26] whose order depends on the complexity of the training set reaching that node. Locally, each node receives a partition of the set received by his parent and it works on such a subset to further partition it for his children nodes. When a leaf node is reached, the final classification is performed. A high-order perceptron is composed by an input layer of M neurons, one for each feature of the input pattern $\mathbf{x} = [x_1, \ldots, x_M]$ to be classified, an intermediate layer where two or more input values are combined in different ways by means of a splicing function (e.g., a multiplication), and an output layer of N elements, one for each problem class. In general, the $k-th$ output y_k of a high-order perceptron is given by (1) where f is a non-linear activation function and w_k are the weights of the connections. The AHNT is automatically grown during the learning phase: it is not needed to choose a-priori the number of nodes and their links. The learning phase consists in feeding the AHNT with vectors defined by all the seven $2^{nd}-$ and $3^{rd}-$ order normalized central moments of the blob image [16], defined as:

$$\eta_{ab} = \frac{\mu_{ab}}{\mu_{00}^{\frac{a+b}{2}+1}} \qquad \forall\, a, b | a+b \in \{2, 3\} \quad (2)$$

where μ_{ab} are the central moments:

$$\mu_{ab} = \sum_{(x,y)\in Blob} (x - x_c)^a (y - y_c)^b \quad (3)$$

and (x_c, y_c) are the coordinates of the barycentre of the blob.

Once the object's class has been assessed a proper identification process is activated as consequence of the type of object currently under analysis.

3.3 Event Analysis

The approaches proposed in the literature can essentially be divided in two categories according to the way events are modelled: implicitly or explicitly. In the former ones, no a-priori knowledge about the domain is provided to the system that automatically identifies common patterns of activity from observed data [17]. The other category includes all the systems that require the user to manually define what constitutes normal and abnormal activity [14]. Implicit modelling makes the system highly adaptable to different scenarios and situations, but inaccurate in detecting specific and complex events. On the other hand, explicit modelling generally yields better results in terms of false alarms and missed alarms, but, of course, this method is not self-adapting as all the knowledge is provided by the operator.

For each target, positional, temporal, and ID information are used for the high-level semantic interpretation of the activities occurring within the monitored scene. Two different types of events have been considered: simple events, characterized by the motion (and behaviour) of a single object (e.g. vehicles, pedestrians, etc. moving in the observed environment) and composite events, characterized by interactions among multiple objects [21]. A composite event is therefore a complex event generated by a set of temporally consecutive simple events or an event composed by multiple moving objects, e.g., a group of people, a queue of cars, etc. A simple event e is defined over a temporal interval $[T^s, T^f]$ and contains a set of features $F = \{f_1, \ldots, f_m\}$ belonging to a given object O_j observed over a sequence of n consecutive frames as:

$$e(T^s, T^f) = \{f_k | f_k \in O_j, k \in [1..m]\} \quad (4)$$

Examples of real features are the ID (i.e., person identification or license plate), the class of the detected object, object trajectories, the average speed, blob shape descriptors, colour histograms, etc.

A composite event is defined over a wide temporal interval as a graph $G(V, E)$ where the set of vertexes V is the set of simple events and the set of edges E is the set of associations (temporal and spatial) between simple events. The events associations are mainly represented by spatial and/or time correlation between simple events.

Once a graph related to a composite event is built, a comparison with the graphs belonging to an event database is performed in order to detect possible threats. The database is set up with a set of explicitly defined complex events sorted in three main categories related to the level of dangerousness: (a) normal events, (b) suspicious events and (c) dangerous events. At the current development, the matching is performed just considering the key nodes that are related to entrance end the exit of the object from the scene (i.e. vehicle/person enters/exits, etc.).

4 Data Transmission

When multiple nodes require to send information to upper nodes through band-limited channels a new problem arises: which is the current capacity of the channel? Moreover, when multiple clients contend for the same resource, it is of paramount importance for each entity to know the share it can exploit to send the data.

In addition, in the case of data that must be further processed at destination, the nodes have also to decide the best representation to convey the information. This aspect is even more important in context of video surveillance. Compressing sensor streams to satisfy the constraints is indeed not always the best choice. The noise that unavoidably is introduced may have dangerous side effects during the processing at the destination.

Bearing in mind these considerations, in the current work we propose an innovative data transmission protocol for a distributed video surveillance system. In particular, we deal with both the way to establish the available bandwidth that each node can dispose and how to determine which is the best representation for the data that has to be sent with respect the priority of the information extracted by each node.

For what concerns the bandwidth assignment the naive solution could consists on splitting the nominal channel bandwidth DR in equal sub-bands such that the data rate DR_i for sensor $i = 1, \ldots, N$ will be $DR_i = \frac{DR}{N}$. This would force the nodes to compress their data to maintain the given constraints regardless the actual available bandwidth and the significance of the content they are sending. Instead, in video security applications, a system that allows to the most important sensors to use a bigger share of the total bandwidth would be more suitable. Therefore, in our solution, we have based the sub-band assignment on two main heuristics. The first consists in estimating on the fly the real available bandwidth (rather than the nominal one) while the second resides on the determination of the sensor priority to decide the bandwidth percentage it deserves. Finally, given these two information a sensor can decide the most appropriate representation of the information to satisfy the constraints.

4.1 Bandwidth Estimation

To determine the available bandwidth for each time instant, the adopted heuristic is based on an easy yet efficient technique based on the computation of the round trip time. For such purposes, at determined time intervals (every $100ms$) an Internet Control Message Protocol ($ICMP$) packet is sent from the decision node to each underlying node that requires to send data. The reason at the root of the decision to use the ICMP round trip time relies on the fact that many of the available communication channels are not symmetric like the Asymmetric Digital Subscriber Line (ADSL), the satellite link etc. This implies different traffic payloads on the two direction thus requiring a mechanism able to measure both. For other type of channels estimating the round trip by the propagation time is enough.

In our solution, by computing the time a packet needs to reach a sensor node and come back to the decision node, a first clue about the congestion level of the network can be inferred. Although it could be adequate, more parameters have been included in the traffic analysis. These are the sum of the data currently received (Total Data Rate TDR), the average number of bits in packet lost PL in the last second and the number of bit in packet received out of order PO in the last second. In

order to come up with a unique global metric function of the traffic inside the network we defined the following function:

$$T = D + \frac{TDR + PL + PO}{DR} \tag{5}$$

where D is the round trip time of the ICMP packet and DR is the nominal data rate of the adopted channel. To have the traffic function T to range from zero to one the round trip time is computed as

$$D = \frac{D_t - \mu}{\sigma} \tag{6}$$

where μ and σ are respectively the mean and the standard deviation of the times computed in the last second and D_t is the current ICMP time.

During real experiments it has been determined that a good value for the traffic parameter is between 0.7 and 0.8. This range represents a trade off between the will of using the entire available bandwidth and necessity of keeping some room for possible bursts coming from the encoded videos. In particular, when active cameras are used, it often happens to have higher data rates when fast movements occur due to the impossibility of detecting the motion vectors.

4.2 Sensor Priority Assessment

For what concerns the decision about the quota of the bandwidth each node can use, we said that in case of a surveillance application the importance of the sensor could be exploited for such purposes. In such a case, we can think to a priority list of sensors whose order is based on the *meaning* of the content acquired by each sensor. A sensor that is acquiring meaningful data for surveillance purposes (event of interest, particular objects, threats, etc.) has to have a higher priority with respect to a sensor that is acquiring normal behaviours. Although it could seem obvious, its assessment is not so naive as it depends on the analysis of the high-level modules within each node. In particular, we need to come up with a significance rate as consequence of object detection, object recognition and event analysis performed by the node. In this plot, if there is no motion occurring inside the monitored area, sending data is almost useless, while sending a video during the tracking of a person's face could represent a high priority.

The proposed solution can be defined on the basis of a lookup table in which each output of the three main modules is associated to a risk index. As can be seen in Fig. 3, the defined table gives a null weight to a motionless situation then it include three classes of risk with increasing priority: a) Object detection, b) Object Recognition and c)Event Understanding. Within each class we have associated different priorities on the basis of the output. It is worth noticing how in each class the undefined or unidentified or unrecognised case has associated the highest index. This is due to the fact that the uncertainty always represent a bigger threat to something that is known. In this point of view, we want a higher priority in order to send more relevant information to the upper nodes in order to relax the uncertainty.

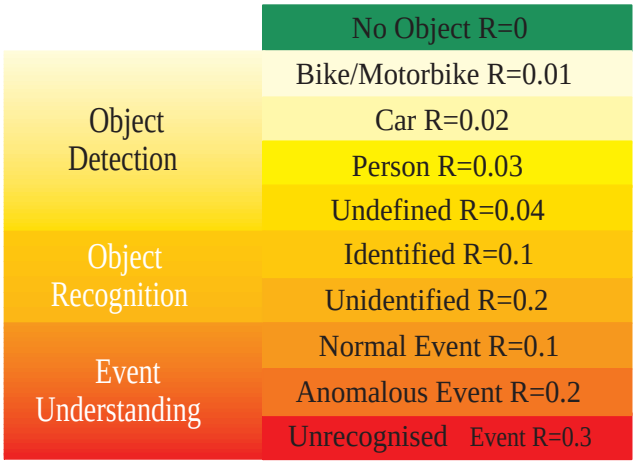

Fig. 3 Look-up table for the priority assignment of each sensor. Every processing step has assigned a risk index R as consequence of its output.

To define the node priority we have associated to each object O_i a set of priority indexes $PI = \{R_{D_i}, R_{R_i}, R_{U_i}\}$ based on the output of the three principal methods executed on them. Let $R_D(O_i)$ be the detection risk in the lookup table for the object O_i and $p_{D_i} = p(O_i)$ the correct classification probability (i.e. output of the neural network taken as a measure about how certain is the class association), then priority index associated to the detection output for the i-th object is defined as

$$R_{D_i} = R_D(O_i) * 1^{(1-p_{D_i})} \tag{7}$$

Notice that, higher the reliability of classification lower the priority risk. This follows the idea to weight more uncertainty than certain outputs. Similar definition can be obtained for $R_{R_i} = R_R(O_i) * 1^{(1-p_{R_i})}$ and $R_{U_i} = R_U(O_i) * 1^{(1-p_{U_i})}$ respectively for the recognition and event understanding tasks with similar definition of the p_{R_i} and p_{U_i} probabilities.

Finally, the total priority index of each node is a weighted combination of the risks computed by the three main modules as follows:

$$Pr = \underset{j,k,l=1..N}{argmax} \left\{ R_{D_j} + R_{R_k} + R_{U_l} \right\} \tag{8}$$

Therefore the priority of a node is given by the sum of the highest risk priorities associated within each class of risk. It is clear how the necessity for a decision node to work on original data compared with the capacity of the network to deliver such information represents a

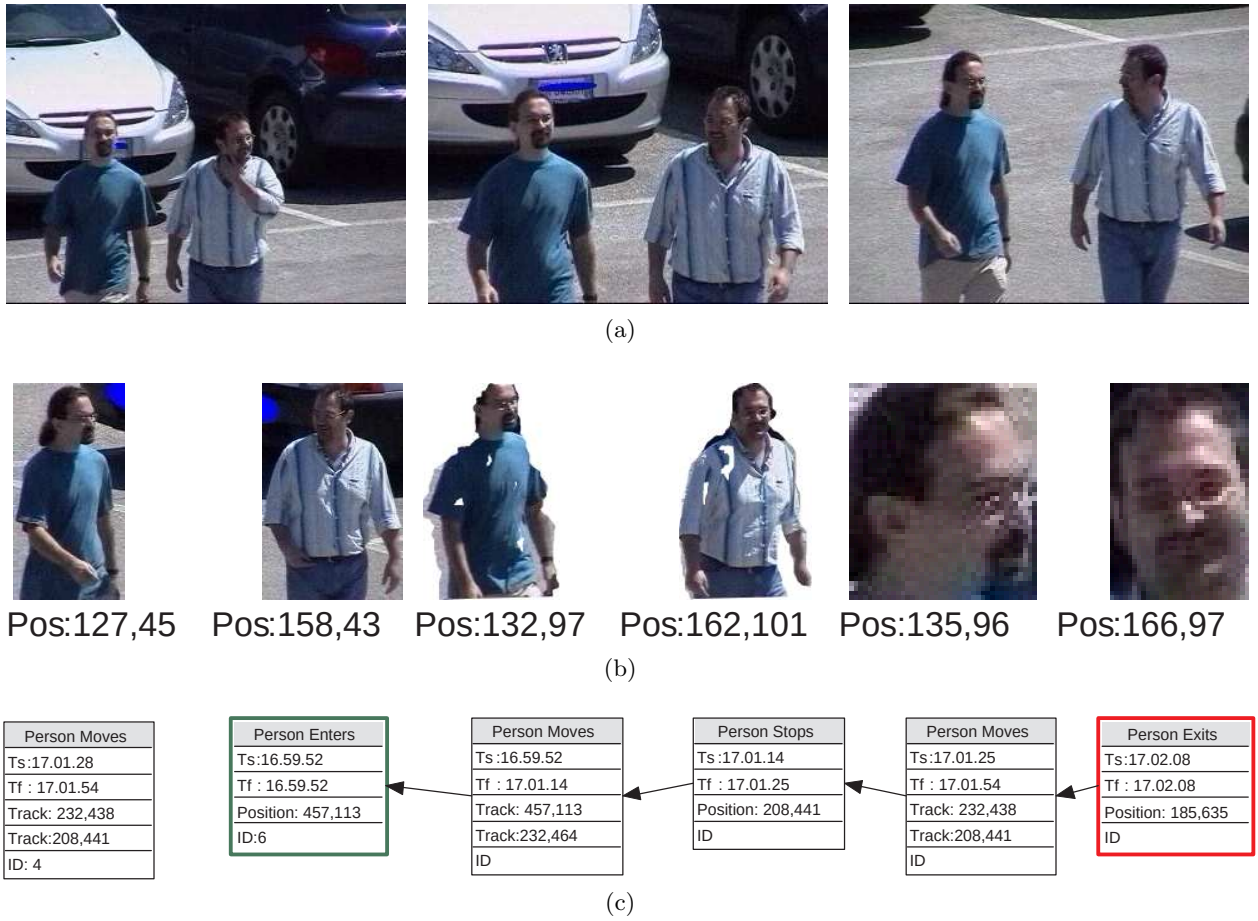

Fig. 4 Example of data transmission. (a) The top row shows a modality when the entire raw data are sent to the decision node. (b) The middle row shows a modality for which reduced data (i.e., bounding box, textured blob, face pattern) are sent together with positional information. (c) At the bottom row, the third modality of transmission is represented: just event information is sent to the decision node (either simple or composite event).

trade-off. Indeed, as the number of sensors acquiring high priority data increases the trust of the decision node on the lower level nodes computation has to increase too. If just few nodes have interesting data, the decision node can request raw data from all of them. Instead, when the case is of many sensors requiring to send useful data, the decision node has to rely on the information processed by underlying nodes.

4.3 Bandwidth Constraints Fulfilment

Once a node receives from its underlying nodes the priorities and after having determined the network traffic, it is responsible to assign the communication parameters to each sensor. Such a process is performed by assigning to each underlying node i a quota Q_i of the current bandwidth as follows:

$$Q_i = \frac{Pr_i}{\sum_{j=1}^{K} Pr_k} \tag{9}$$

where Pr_i is the priority received by node i and K is the number of nodes requiring to transmit to the decision node. Once the DN has decided the share for node i this has the duty to satisfy the given constraint. To achieve such an objective he can decide among three types of transmission modalities:

- Raw data
- Raw data and control information
- Descriptive information

In the first case, the node considers that the assigned bandwidth is sufficient for the transmission of the full video as close as possible to the data acquired by the sensor. The only action that is required is the tuning of the compression parameters. In this context, the node continuously tunes fundamental parameters like the compression level and the frame rate to meet the determined bandwidth constraints.

If the assigned bandwidth does not allow to send such a quantity of information, the node decides to extract useful control information to provide just a small portion

of the acquired data. This can be the case of sending the bounding box of a tracked object together with data like current position, trajectory performed so far, etc. In this case, the amount of data sent drops dramatically especially in the case of data coming from active camera (static macroblocks can be easily encoded by the MPEG-4 algorithm).

Finally, if the amount of sensors requiring to send useful data increases to a critical level, the node is required to use a very limited quota of the bandwidth. Therefore, it has no other choice than choosing the lightest bit rate transmission. In this case, only information extracted by the underlying nodes are requested to be sent without requiring the forward of original data.

In Fig. 4 an example of data sent within the three possible modalities is shown.

5 Experimental Results

The proposed system has been extensively tested in all its parts by adopting a strategy that follows an incremental complexity of the events. In addition, all the modules of the system have been first singularly tested then their behaviours have been altogether checked.

To test the communication system two types of networks have been considered. A first one consists in a satellite link to transmit data acquired from sensors deployed in a remote environment to monitor a restricted access area. The second consists in a 802.11g wireless network installed at the university building where a parking lot has been chosen as test bed site.

For what concerns the object classification, an AHNT has been trained to classify the objects into three main classes: cars, cycles and pedestrians.

Security Tasks

Concerning the event analysis, the system has been tested on sequences taken in a parking lot and showing different possible cases. Each sequence was manually labelled with ground truth data, identifying which kind of simple behaviour (normal, suspicious or dangerous) was happening in the sequences. A feed forward neural network has been trained with a back propagation technique on about 50 patterns representing normal events, 30 patterns for suspicious events and 20 for dangerous events, for both vehicles and pedestrians. The normal events are further processed by the composite event detector, in order to spatially and/or temporally correlate simple events. Finally, from a match with the event database a definitive classification in normal or anomalous behaviours is performed. Since such a classification relies on the similarity between two graph of events, it is possible to have an unrecognised behaviour when the current event does not match with any event of the database.

In Fig. 5 two different first level nodes A and B, looking at different areas, can detect and recognize simple events (vehicle/person related) and from them build composite events. In this example node A (see Figure 5(a)) detects a car entering a parking lot to park in a free spot. Afterwards, a person is getting out of the car to walk out of the scene. Meanwhile, node B (see Fig. 5(b)) starts to detect a person entering its area and moving toward a parked vehicle. The temporal and spatial correlation between the exit from the scene of the person and the initial motion of the car, trigger the association of the two events. This yields to the definition of a complex event related to a person getting into the car.

Both events, singularly taken, seem to be normal events. Therefore, without information sharing between these two nodes no alarms would raise. Instead, if we allow a decision node to process the two videos, we obtain an interesting result. The association between the two persons (the one walking out the field of view of sensor A and the one entering the field of view of sensor B) is made. In this case, the node puts together a new composite event (see Fig. 5(c)). The system recognized a person parking a car and getting out of the scene aboard a second car. This is one of the dangerous events that have been loaded into the event database as a dangerous behaviour (i.e. person parking a car bomb near a target).

Data Transmission

To test the communication capability of the system we have conducted two main tests. A first one consisted in sending a video through a satellite link by simulating traffic to see how the system responds to casual congestion not due to data flowing from the sensors. In Fig. 6 the results obtained in such a context are presented.

It is worth noticing how the system, as consequence of a traffic peak, requires the sensor to drastically reduce the quality and the frame rate to maintain active the transmission. Once the network traffic reduces, the system raises both sensor parameters to obtain better video quality.

To test the efficacy of the data transmission protocol we sent two streams with different priorities. In particular, the highest priority $Pr = 1$ (corresponding to an unknown object performing an unrecognised event) has been assigned to an active camera while a low priority $Pr = 0.3$ (corresponding to a known person performing a common behaviour) has been assigned to a static camera. As can be seen from the charts depicted in Fig. 7, the bandwidth assigned to the active camera is much bigger than the one assigned to the static camera. As result we can notice how both quality and frame parameters are set to the maximum for the highest priority sensor. Instead, when the network traffic increases the system correctly tunes both quality and compression parameters of the static sensor to maintain the traffic metric within the desired range. In particular, it is interesting to notice how both the compression quality and the frame rate concerning the stream with lower priority does not increase before the traffic falls below the value

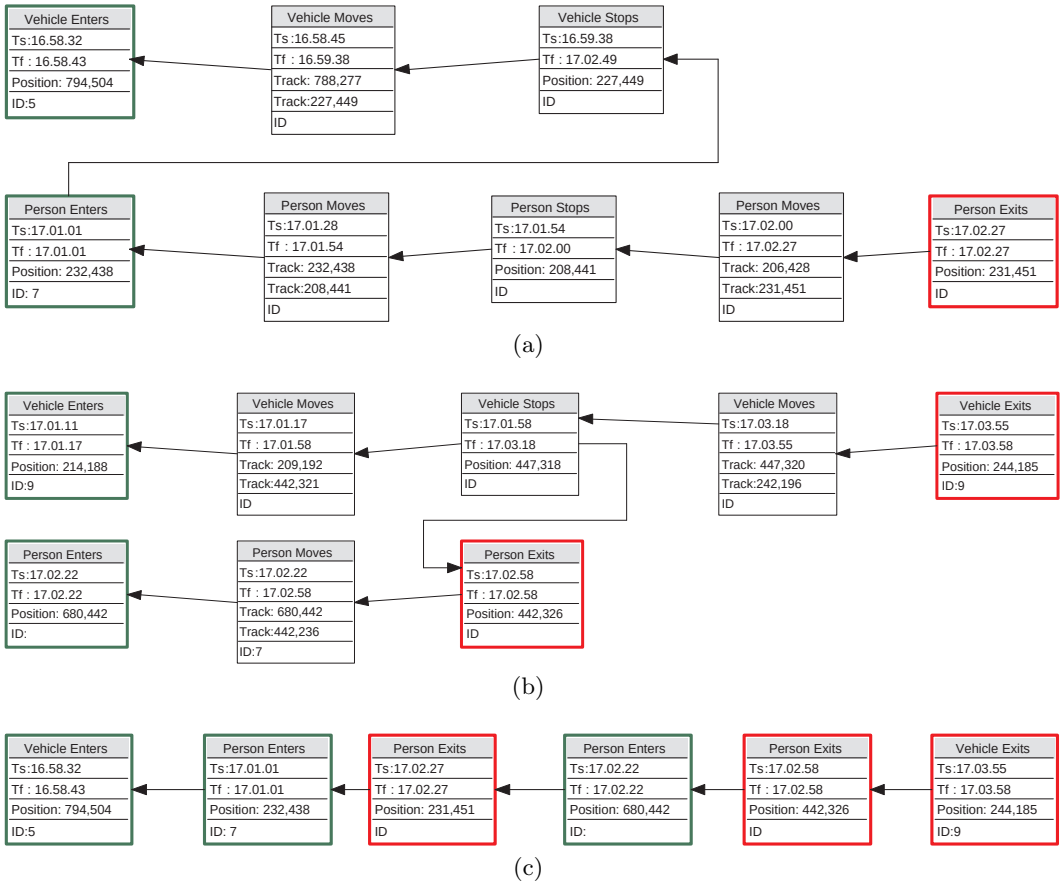

Fig. 5 Example of composite event detection (arrows represent event association). In (a) the event detection performed by a first level node A is shown. In (b) a similar computation performed by a node B (monitoring a different area) is shown. In (c) it is shown a graph of a more complex event recognition performed by a decision node after having received data from nodes A and B.

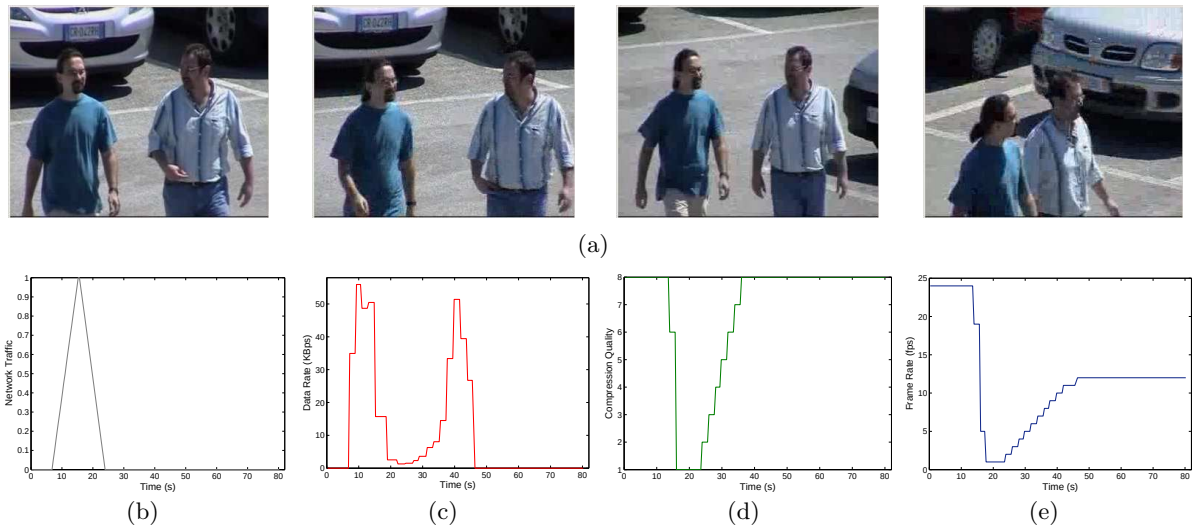

Fig. 6 Experimental testing of the automatic adaptation of the sensor parameters as consequence of the network traffic estimation. (a) Four of the received frames representing respectively from left to right high quality, medium quality,low quality and again medium quality compressed frames. (b) Simulation of the traffic peak. (c) Resulted data rate as consequence of the selected compression quality (d) and of the frame rate (e).

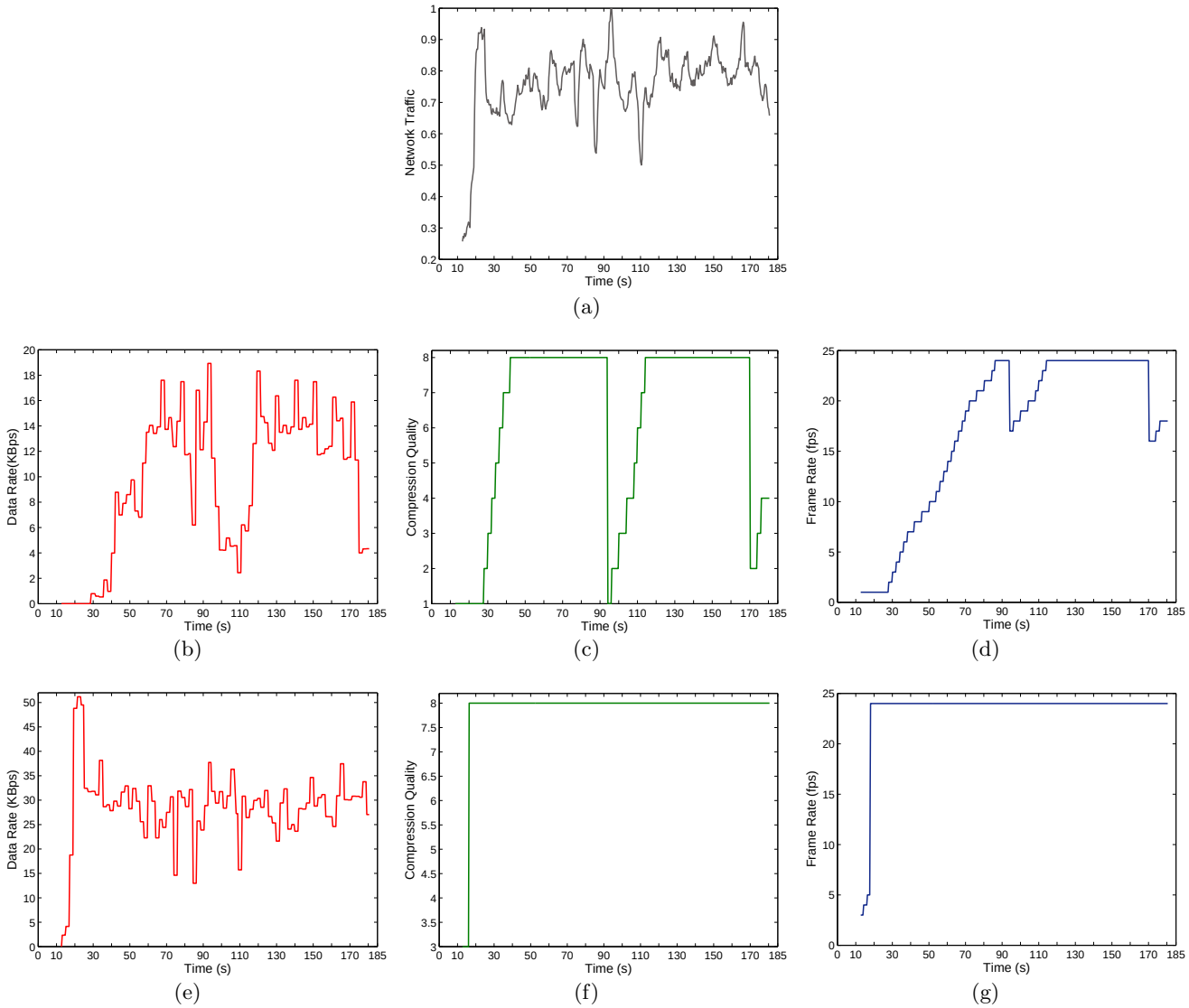

Fig. 7 Example of automatic transmission parameters configuration. The aim of the experiment was to analyse the capability of the system to assign a bigger subband to sensor B (last row) then to sensor A (middle row). As can be seen the sensor B corresponding to an active camera with $PR = 1$ received enough bandwidth to send its data with the highest frame rate and compression quality. Instead, it is worth noticing how as the network traffic increases the system automatically decreases both compression quality and the frame rate of sensor B which has a priority $Pr = 0.3$.

0.8. This is the result of the will to maintain the traffic parameter in the range $[0.7 - 0.8]$. Once such a value is obtained, the node increases both parameters as soon as it receives the authorization to use a bigger quota of the total bandwidth.

A last experiment to test the improvements that the proposed data transmission brings to the security tasks has been finally conducted. In this experiments, we have simulated a situation in which a node is monitoring a zone of the university parking lot and its supervisor node is monitoring a nearby area. To let us to make comparisons with single unconnected nodes we used the videos adopted to test the security tasks (see Fig. 5). In particular, we sent the video of node A through a 802.11g wireless network to a decision node processing the video of node B. The channel adopted was characterised by a bandwidth limited to 1024Kbps, the spatial resolution of the frames was 320×240 with 24 bit colour resolution. In addition, with respect to the previous experiments we have increased the number of intervals for the compression regulation in order to achieve a better quality for the transmission.

Initially we tested the proposed solution which was able to react to the traffic flow of the network by tuning the compression and frame-rate parameters of video A (see Fig.8). It is worth noticing how the number of packets lost is kept to zero by the tuning algorithm (see Fig. 8(e)). In addition, see Fig. 8(b), the total data rate

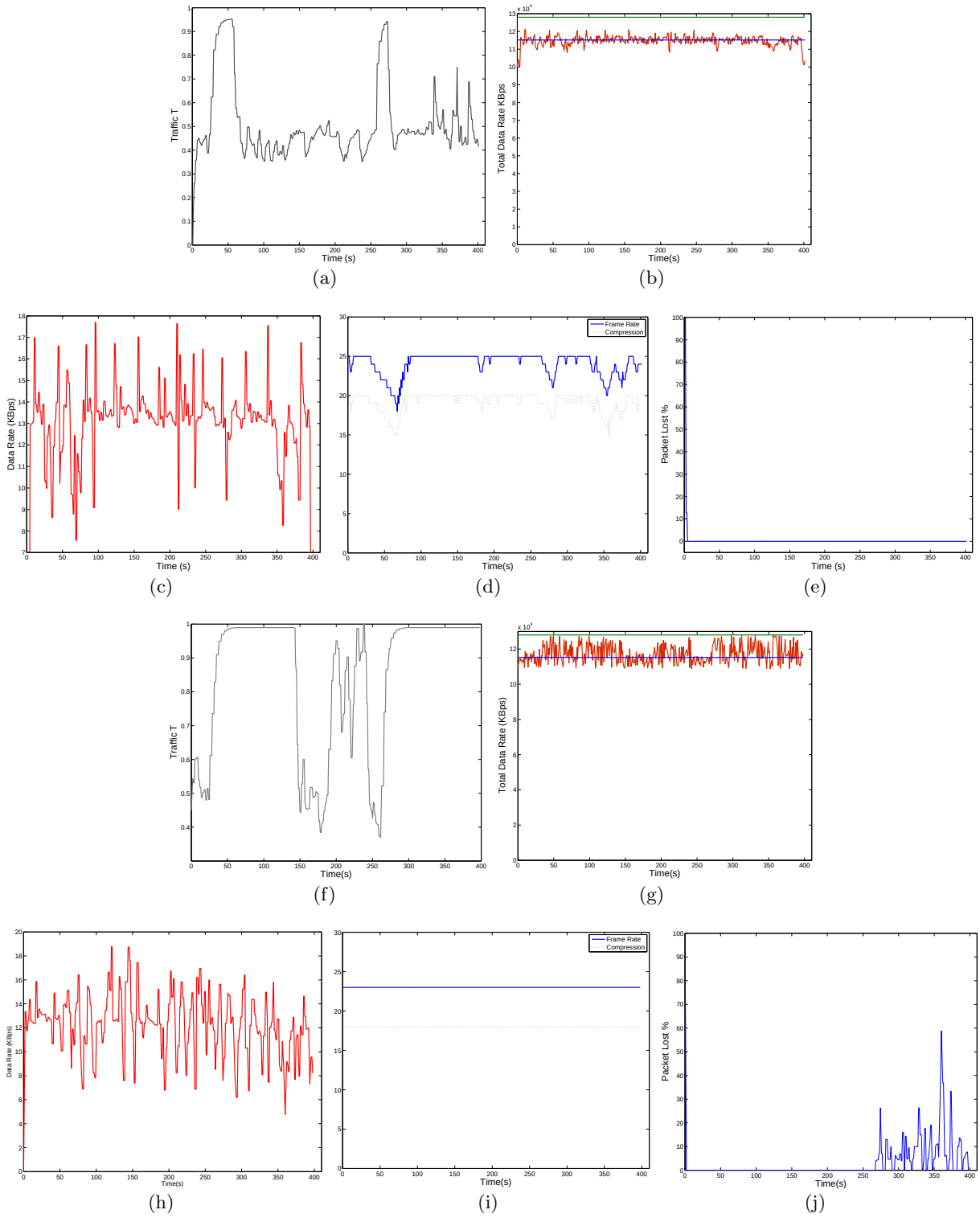

Fig. 8 Example of the effectiveness of the proposed solution compared to the performance of a common data transmission protocol. From chart (a) to (e) data concerning respectively the estimated traffic, the total data rate, the video data rate, compression and frame rate parameters and the percentage of packet lost during transmission of video A with the proposed data transmission algorithm are plotted. Charts (f)-(j) show the same data computed by using a common data transmission protocol without the tuning of the parameters.

is maintained close to the 85% of the nominal data rate (the top continuos line represents the nominal data rate while the lower line it 85%). Such a result allowed the decision node to process entirely the information acquired by node A. Hence, the security tasks executed on the received video has been able to detect the same events of the previous experiment. These events, once correlated by the decision node, yielded the same complex event recognition.

As second step, we computed the mean values of compression level and frame rate obtained by the automatic tuning. Therefore, these values (18 and 23 respectively for the compression and the frame rate) has been set for video A. Then, we have switched off the data transmission algorithm and sent the compressed video through the same network. As result, we have identified a major request of bandwidth (see Fig. 8(g)) that has involved a considerable increment of packet lost (see Fig.8(j)). In particular, we obtained a peak rate of about 60% of packet lost.

The consequence of such number of data lost has been the impossibility of the decision node to process the video. Important data has been definitely lost. The major effect has been an unsuccessful detection of the simple events for video A. It is worth noticing, see Fig. 9, how the output of the detection module (Fig. 9(b)) on a corrupted frame 9(a),received during a phase of packet lost, is unusable.

In addition, also the correlation between events generated by video A with those generated by video B has been impossible thus avoiding the recognition of a possible threat within the monitored area.

6 Conclusions

The rapid development of sensors, network communications and computational systems forces the researchers working in the field of advanced surveillance systems to design and develop innovative solutions for building "intelligent" systems able to support human operators in the task of making complex decisions. As both civil and military applications require to monitor large and complex environments, the new generation of video surveillance systems needs to use a large numbers of heterogeneous sensors fully interconnected. In this paper, a system architecture with distributed intelligence over multiple processing units and with an innovative distributed communication over multiple heterogeneous channels (wireless, satellite, local IP networks, etc.) has been proposed. A new real-time technique able to tune the most important video transmission parameters like the frame rate and the spatial/color resolution on the basis of the available bandwidth has been presented. Experimental results have shown the good performances of the main system modules and the robustness of the communication procedure. Two kinds of networks have been considered: a satellite link and a 802.11g wireless network installed at the University building. It has been shown also that the threat recognition capability of the distributed system without the proposed communication protocol can show important flaws with respect to the response capability. Indeed, threats that can be detected by exploiting the proposed communication system would remain undetected with a normal communication system.

Acknowledgments

This work was partially supported by the Italian Ministry of University and Scientific Research within the framework of the project "Ambient Intelligence: event analysis, sensor reconfiguration and multimodal interfaces"(2007-2008) and by the European FP6 Project HAMLeT "Hazardous Material Localisation & Person Tracking" (SEC6-SA-204400).

References

1. Bhandarkar, S., Luo, X.: Fast and robust background updating for real-time traffic surveillance and monitoring. In: IEEE International Conference on Computer Vision and Pattern Recognition, vol. 3, pp. 55–59. San Diego, CA, USA (2005)
2. Bjontegaard, G., Lillevold, K., Danielsen, R.: A comparison of different coding formats for digital coding of video using mpeg-2. IEEE Transactions on Image Processing **5**(8), 1271–1276 (1996)
3. Collins, R., Lipton, A., Fujiyoshi, H., Kanade, T.: Algorithms for cooperative multisensor surveillance. Proceedings of the IEEE **89**, 1456–1477 (2001)
4. Collins, R., Lipton, A., Kanade, T., Fujiyoshi, H., Tsin, D.D.Y., Tolliver, D., Enomoto, N., Hasegawa, O.: A system for video surveillance and monitoring. Tech. Rep. CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2000)
5. Fan, J., Yau, D., Elmagarmid, A., Aref, W.: Automatic image segmentation by integrating color-edge extractionand seeded region growing. IEEE Transactions on Image Processing **10**(10), 1454–1466 (2001)
6. Foresti, G., Dolso, T.: Adaptive high-order neural trees for pattern recognitionadaptive high-order neural trees for pattern recognition. IEEE Transactions on System, Man and Cybernetics Part B **34**(2), 988–996 (2004)
7. Foresti, G., Micheloni, C., Snidaro, L., Remagnino, P., Ellis, T.: Advanced image and video processing in active video-based surveillancesystems. IEEE Signal Processing Magazine **22**(2), 25–37 (2005)
8. Foresti, G., Regazzoni, C., Visvanathan, R.: Scanning the issue technology - special issue on video communications,processing and understanding for third generation surveillance systems". Proceedings of the IEEE **89**(10), 1355–1367 (2001)
9. Haritaoglu, S., Harwood, D., Davis, L.: W^4: Real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8), 809–830 (2000)
10. Hata, T., Kuwahara, N., Nozawa, T., Schwenke, D., Vetro, A.: Surveillance system with object-aware video transcoder. In: International Workshop on Multimedia Signal Processing. Shangai, Cina (2005)

(a) (b)

Fig. 9 Fig. (a) is a frame of video A as it has been reconstructed by the decision node during a phase of packets lost (see Fig. 8(j)). It is interesting to notice how the developed low level module for the motion detection fails to detect the real object (i.e. a person). Indeed, the resulting image (Fig. (b)) has the majority of the pixels classified as belonging to moving objects(white pixels).

11. ISO/IEC: Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 2, 11172-2 edn. (1993)
12. ISO/IEC JTC1: Coding f audio-Visual Objects -Part2: Visual", iso/iec 14 496-2 edn. (1999)
13. ITU-T: Video Codec for Audiovisual Services at px 64 Kbit/s Version 1, itu-t reccomendation h.261 edn. (1990)
14. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Transactions on Pattern Analysis Machine Intelligence **22**(8), 852–872 (2000)
15. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A new method for gray-level picture thresholding using the entropy of the histogram. Graphical Models and Image Processing **29**, 273–285 (1985)
16. Liao, S., Pawlak, M.: On image analysis by moments. IEEE Transactions on Pattern Analysis and Machine Intelligence **18**, 254–266 (1996)
17. Makris, D., Ellis, T.: Learning semantic scene models from observing activity in visual surveillance. IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics **35**(3), 397–408 (2005)
18. Micheloni, C., Foresti, G., Snidaro, L.: A network of cooperative cameras for visual-surveillance. IEE Visual, Image & Signal Processing **152**(2), 205–212 (2005)
19. Otsu, N.: A threshold selection method from gray level histograms. IEEE Transactions on System, Man and Cybernetics **SMC-9**, 62–66 (1979)
20. Petersen, J.: Understanding Surveillance Tehnologies. CRC Pess (2001)
21. Piciarelli, C., Foresti, G.: Event recognition by dynamic trajectory analysis and prediction. In: IEE Image for Crime Detection and Prevention. Savoy Place, London UK (2005)
22. Ridler, T.W., Calvard, S.: Picture thresholding using an iterative selection method. IEEE Transactions on System, Man and Cybernetics **SMC-8**, 630–632 (1978)
23. Rosin, P.: Thresholding for change detection. In: Proceedings of IEEE International Conference on Computer Vision, pp. 274–279. Bombay India (1998)
24. Snidaro, L., Foresti, G.: Real-time thresholding with Euler numbers. Pattern Recognition Letters **24**(9-10), 1533–1544 (2003)
25. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Pattern Analysis and Machine Intelligence **22**(8), 747–757 (2000)
26. Thimm, G., Fiesler, E.: High-order and multilayer perceptron initialization. IEEE Transactions on Neural Networks **8**(2), 349–359 (1997)
27. Wiegand, T., Scwarz, H., Joch, A., Kossentini, F., Sullivan, G.: Rate-constrained coder control and comparison of video coding standard. IEEE Transactions on Circuits and Systems for Video Technology **13**(7), 7 (2003)
28. Zitova, B., Flusser, J.: Image registration methods: a survey. Image and Vision Computing **21**, 977–1000 (2003)
29. Z.Tao, Nevatia, R.: Tracking multiple humans in complex situations. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), pp.1208–1221 (2004)