The Journal of SuperComputing

Real-Time Wireless Multisensory Smart Surveillance with 3D-HEVC Streams for Internet-of-Things (IoT)

George Kokkonis, Kostas E. Psannis, Manos Roumeliotis, Dan Schonfeld

Abstract This paper presents the design of a novel, real-time, wireless, multisensory, smart surveillance system with 3D-HEVC features. The proposed high-level system architecture of the surveillance system is analyzed. The advantages of HEVC encoding are presented. Methods for synchronization between multiple streams are presented. Available wireless standards are presented and compared. A network-adaptive transmission protocol for a reliable, real-time, multisensory surveillance system is proposed. Adaptive Packet Frame Grouping (APFG) and adaptive quantization are deployed in order to maximize the Quality-of-Experience (QoE). Measurements of the proposed protocol have been shown to provide superior results compared to existing transport protocols.

Keywords Smart surveillance, Internet-of-Things (IoT), HEVC, depth cameras, real-time transport protocols, wireless transmission.

1 Introduction

The Internet has revolutionized the computer applications and communications. From the simple transfer of text messages we have moved to the Web 2, the Machine to Machine (M2M) communications and the transfer of real-time audio, video and other multisensory data, all together called supermedia. The new evolution stage of the Internet is the Internet-of-Things (IoT). Internet of Things semantically means a world-wide network of interconnected object uniquely addressable, based on standard communication protocols [1]. Person-to-person, person-to-object and object-to-object communication is now thriving. Humans, sensors, actuators, and smart objects are exchanging vast, real-time information.

One main sector of the IoT is the efficient encoding and transferring of video streams over the Internet. The new High Efficiency Video Encoding (HEVC) offers more than 50% improvement in video compression over its predecessor H.264 Advanced Video Coding standard, with the same image quality, at the expense of the increased computational complexity [2]. The 50% improvement of video compression corresponds to 50% bit rate reduction for video transferring over the IoT. This reduction is crucial, especially in real-time video monitoring for smart surveillance systems. Smart surveillance, is the use of automatic video analysis technologies in video surveillance applications. Smart surveillance systems often deal with real-time monitoring of persistent and transient objects/people within a specific environment. They use image processing algorithms for detection, tracking, and understanding of moving objects/people of interest in dynamic scenes. The main stages of processing in an intelligent visual surveillance system are: moving object detection and recognition, tracking, behavioural analysis and retrieval [3].

The smart surveillance systems are divided in two main categories, depending on where the video analysis is made. The first category includes cameras with processing capabilities. The detection of the event and the storage of the event are made autonomously by the camera. Such cameras are called smart cameras [4]. In the second category, the video stream is transferred through the network in datacenters where the video analysis is made. Large research projects on this category are the Visual Surveillance and Monitoring (VSAM) [5], the Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval (ADVISOR) [6], and the Smart Surveillance System of IBM [7]. In the Third Generation Surveillance Systems (3GSS) [8] a mixture of these categories

G. Kokkonis, K. E. Psannis, M. Roumeliotis
Department of Applied Informatics,
University of Macedonia,
156 Egnatia Street,Thessaloniki 54006 , Greece.
E-mail: {gkokkonis, kpsannis, manos} @uom.gr

D. Schonfeld
Department of Electrical & Computer Engineering
University of Illinois at Chicago
Chicago, IL 60607-7053
E-mail: dans@uic.edu

is made, both smart cameras and large datacenters are being used. The video analysis often includes image enhancement [9], motion detection [10], object tracking [11], and behavior understanding [12].

Smart surveillance systems can be enhanced with the help of 3D monitoring technologies. The 3D camera, also called time-of-flight or depth camera, emits modulated infrared light and measures the time the infrared signal takes to travel from the camera to the object and back again: the elapsed time is called "time of flight".

As we are heading towards to the 5th generation of wireless/ mobile broadband networks, numerous devices and networks are interconnected. The Internet of Things (IOT) is becoming a reality. Person-to-person, person-to-object, and object-to-object are exchanging continuously massive real-time supermedia data. The efficient transmission of this data is of great importance for the smart surveillance systems. The scope of this paper is to propose a transport protocol for efficient delivery of supermedia data from smart surveillance systems. Flow/congestion control algorithms and synchronization techniques should be enforced in order to maximize efficiency of a smart surveillance system.

The rest of the paper is organized as follows: Section 2 presents a high-level system architecture for smart surveillance systems. Section 3 outlines the characteristics of an HEVC – 3D depth video monitoring. Section 4 propose a synchronization algorithm for intra and inter media synchronization between RGB video, Depth video, audio and other multisensory data. Section 5 presents the wireless network infrastructure of the surveillance system. Section 6 analyzes the proposed protocol for transferring multisensory data streams through the wireless IoT networks. Section 7 presents a case study for transferring a wireless multisensory 3D - HEVC stream over wireless networks. Finally section 8 concludes this paper.

2 Smart Surveillance Systems

As the number of threats of burglary, robbery and terrorist activities increase, surveillance systems becomes a necessity. The traditional methods for monitoring with the use of CCTV cameras are now evolving to smart surveillance systems. Threat detection and video analysis is automatically made by information systems. Instant notification and alerts produced by smart surveillance systems reinforce public safety and security.

Smart surveillance data collected by cameras and sensors can also provide valuable decision support analysis to organizations. Intelligent analytics enhance business intelligence. The automatic video analytics help organizations to seize opportunities or revise policies.

Smart surveillance systems form a very important sector of video monitoring, video transmission and video analysis. Automated recognition of individuals and objects, pre-determined traits and risks lies at the basis of smart surveillance systems. Some of the innumerable applications of the smart surveillance systems are:

- Unattended surveillance for security reasons
- Automated inspection for quality assurance
- Defect detection and dimensional gauging
- Non contact measurements
- Part sorting and identification
- Code reading and verification
- Robot guidance and automated picking
- Biometric recognition and access control
- Object detection and tracking
- Environment mapping

Apart from video, a smart surveillance system can monitor many more sensory data such as:

- Audio
- Temperature
- Humidity
- Acceleration
- Luminosity
- Pressure
- Chemical Analysis
- “Time of Flight” depth imaging
- Radiation
- Ultrasonic waves
- Motion

All these sensory information is being recording by sensors which are connected to the surveillance system via the net. The Internet of Things is the network which enables these sensors to collect and exchange data through the existing network infrastructure. It allows sensors to be identified and controlled remotely. The IoT improves efficiency, effectiveness. The transport protocols that have already been proposed for the IoT are the MQTT, the XMPP, the COAP, and the 6LOWPAN [13]. Energy consumption, safety and scalability are the main concerns of these protocols.

A proposed high-level system architecture of a multisensory smart Surveillance system is depicted in Figure 1.

In the proposed system architecture three are the main channels that are transferring information between the monitored environment and the data center.

The first channel is the **Video Channel**. It transfers visual information from the monitored environment to the data center. In order to lower the bandwidth needed for this transmission an HEVC encoder is used to lower the bitrate, while keeping the image quality as good as possible. Network-adaptive congestion control algorithms can be used to absorb all the available bandwidth without forcing the network to congestion [14]. The Real-time Transport Protocol (RTP) is often used to transfer information in this transmission.

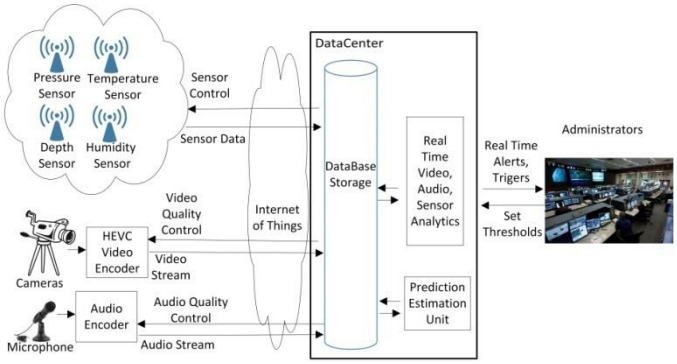

Fig. 1. High-level system architecture of a multisensory smart surveillance system.

Another important channel that is often used in smart surveillance systems is the **Audio Channel**. Advanced Audio Coding (AAC) [15] is usually used for the audio encoding, which is a lossy data compression technique for digital audio. The RTP transport protocol is also used for this channel.

The **Multisensory Channel** is the third data transferring channel. It carries all the sensory data from the monitored environment to the data center. A real-time, network-adaptive, proposed by the authors, protocol is used for the transport of this information. It is often a channel with low-bandwidth requirements but with high update rate frequencies. Moreover, algorithms that enhance the reliable transfer of data should be applied.

The **Data Center** stores all the information from the remote monitored environment. It is often located away from the monitored environment at a secure location, connected to the Internet. All the data stored there are analyzed and real-time alerts are triggered, based on the thresholds set by the surveillance administrators. The data center also includes a Prediction and Estimation Unit (PEU). Its main functionality is to perform prediction algorithms on historical database records. Future potential risks can be avoided with the PEU. Data mining algorithms for statistical model building are also applicable.

3 High Efficiency Video Coding (HEVC) with 3D Features

3.1 HEVC Video Encoding

The High Efficiency Video Coding (HEVC) is ideal for real-time video transferring through the Internet. The HEVC standard achieves more than 50 % improvement in video compression and bitrate over the existing H.264 Advanced Video Coding standard, for the same image quality [16]. The expense of this achievement is the increased computational complexity, which is compensated from the improvement of the computational power of the active equipment. The 50% reduction of the bitrate together with the improvement of Internet network conditions made feasible the real-time streaming of high- and ultra-high-definition video over the Internet. In order to compare the bitrate between the HEVC and its predecessor H.264, the video sample mobile_cif YUV series with resolution 352 × 288, refresh rate 24 Hz and duration 10 sec [17] was encoded with both encoders. The encoding time, the bitrate and the PSNR are depicted in Table 1. For the HEVC encoding, the HEVC Test Model HM 16.2 has been used. For the H.264 encoding the H.264/14496-10 AVC - JM 18 encoder was used. A PC Intel core i3 2100 at 3.1 GHz with 4 GB RAM was used for the encoding process.

Table 1. Computational Cost of HEVC

	COMPUTATIONAL COST of HEVC (for 10 sec video)				
Inter Prediction	***QP***	***RAM (MB)***	***Data Rate (Kbps)***	***Encoding Time (sec)***	***PSNR***
Intra-Only	32	20.9	3514	341	33.18
Low-Delay	32	47.8	269	1209	31.18
Random Access	32	64.5	290	807	31.66
Intra-Only	27	21.1	5266	374	37.26
Low-Delay	27	47.4	641	1491	34.60
Random Access	27	64.7	584	1012	34.76

Table 2. Computational Cost of H.264/AVC

	COMPUTATIONAL COST of H.264/AVC (for 10 sec video)				
Inter Prediction	QP	***RAM (MB)***	***Data Rate (Kbps)***	***Encoding Time (sec)***	***PSNR***
Intra-Only	32	32.3	3948	1191	33.99
Low-Delay	32	95.9	652	2983	33.37
Random Access	32	111.8	270	9528	32.58
Intra-Only	27	32.3	5974	1416	37.61
Low-Delay	27	80.3	1490	2702	36.70
Random Access	27	112.7	644	6487	35.43

As smart video surveillance focuses in real-time video transmission, the encoding time and the data rate are very important factors for the selection of the inter prediction mode. Between the Intra-only and the Low-delay mode of the HEVC encoder, the second is preferred as it presents much lower data rates. Between the Low-Delay and the Random-Access mode, again the Low-Delay mode is preferred because the encoder doesn't have to wait the whole GoP to appear in order to start the encoding process [18].

Comparing Tables 1 and 2 for the Low Delay mode, it is obvious that the data rate with the HEVC encoder is reduced by 58.74 % for Quantization Parameter (QP) =32 and 56.91% for QP=27. This reduction indicates HEVC as perfect encoder for real-time video transmission.

3.2 Depth Video

Apart from the ordinary video, another interesting factor that can be used in the smart video surveillance is the 3D-Depth cameras. These cameras measure the distance between the depth camera and an object with the help of modulated infrared light. Given that the speed of light is known, the "time of flight" that takes the light to travel from the camera to the scattered object and return back reveals the distance between the camera and the object. If this distance is measured for every pixel of the video, the ordinary video is transformed to a 3D-depth video. The depth video can enhance smart surveillance at motion surveillance, object tracking, acceleration and speed monitoring, and 3D-environment mapping. “Time of Flight” monitoring is invariant to visible lighting and weather conditions. The first widely used depth camera was the Microsoft Kinect camera [19] for the Xbox. It produces an 11-bit 640 × 480 depth image at 30 frames-per-second. Its precision is inversely proportional to the distance, and is decreasing from 1 cm at two meters to 10 cm at six meters. If no compression technique is enforced, the data rate for an 11-bit 640 × 480 depth image at 30 fps is 3379,2 Kbps, which is a high data rate for real-time transmission over the Internet. Therefore, intra- and inter-frame compression techniques should be applied to the depth stream in order to limit the data rate.

A sample of an RGB-Depth image is depicted at the right side of Figure 2. The depth image, at the middle was produced by a Microsoft Kinect camera [20]. It is obvious that most of the objects of the left RGB picture can easily be distinguished and surveilled in the processed RGB-D image.

Fig. 2. Output from an RGB camera (left), preprocessed depth (center) and processed RGB (right) [20].

4 Synchronization of HEVC Streams with Depth Video and Multisensory Data

The video camera, the depth camera, the audio and all the other sensors produce frames at a different rate, which means that these frames are loosely coupled. In order these frames to be synchronized, a synchronization algorithm should be enforced. The proposed synchronization algorithm is the enhanced Virtual-Time Rendering (VTR) media synchronization algorithm [21].

Enhanced VTR enforces both intra and inter synchronization in both streams. The intra synchronization tries to keep the outputs media Units (MUs) to the destination, at the same intervals as the generation ones at the source. The inter synchronization reconstructs the original temporal relations between the MUs of different streams.

In order the enhanced VTR to be enforced, one of the multisensory streams should be defined as the master stream. As the smart surveillance mostly depends on the visual sense, the video stream is often selected as the master stream.

4.1 Intra Synchronization

The VTR algorithm enforces the intra synchronization for each stream separately. Each Media Unit (MU) that is sent from the source should be an output at the destination at an ideal output time x_n. If G_n, A_n, and O_n are the generation time, the arrival time and the output time of *n-th* MU respectively, then the ideal output time x_n is given by the Eq. (1) and (2)

$$x_1 = \begin{cases} A_1 + J_{max}, & \text{if } A_1 + J_{max} - G_1 \le D_{max} \\ G_1 + D_{max}, & \text{otherwise} \end{cases} \quad (1)$$

$$x_n = x_1 + (G_n - G_1) \quad (n \ge 2), \quad (2)$$

where D_{max} is the maximum allowable delay and J_{max} is the maximum allowable jitter.

Eq. 1 determines the output time of the first packet x_1, while Eq. 2 sets the outputs times $x_n (n \ge 2)$ at the same intervals as the generation ones at the source.

If jitter is larger than J_{max}, the ideal output time x_n should be changed to *target output time* t_n.

$$t_1 = x_1, \quad \Delta S_1 = 0 \quad (3)$$

$$t_n = x_n + \sum_{i=1}^{n-1} \Delta S_i \quad (n \ge 2) \quad (4)$$

$$t_n^* = t_n + \Delta S_n \quad (n \ge 2) \quad (5)$$

Where t_n^* and ΔS_n is the *modified target output time* and the *slide time*, respectively.

If the *n-th* packet arrives earlier (A_n) than the *modified target output time* t_n^*, the *scheduled output time* $d_n (n \ge 2)$ is equal to the *modified target output time* t_n^*. If the *n-th* packet arrives later than the t_n^*, it is outputted as soon as it arrives, Eq. 6.

$$d_n = \begin{cases} t_n^*, & if\ A_n \le t_n \\ A_n, & otherwise \end{cases} \quad (6)$$

When multiple MUs have the same scheduled output time, the MU which has the largest sequence number is outputted

and the other MUs are skipped.

The *slide time* ΔS_n is calculated as follows:

(a) Virtual-Time Expansion

If $d_n - t_n > T_{h2} > 0$, then $\Delta S_n = d_n - t_n$, where T_{h2} is a threshold which decides whether the target output time should be delayed or not.

(b) Virtual-Time Contraction

If $A_n \leq t_n$, then $d_n = max(t_n - r, x_n, A_n)$ and $\Delta S_n = -min(r, \sum_{i=1}^{n-1} \Delta S_i)$ when $t_n - T_n > D_{max}$, or if a period of time ($T_{nodelay}$) has elapsed since the last late arrival or the last virtual-time contraction, where r is a positive constant. If $d_n \leq O_m$ $(n > m)$, the *n-th* MU is skipped.

4.2 Inter Synchronization

As soon as the intra synchronization for all the data streams is started, the inter-synchronization among all the streams should be enforced. The inter-synchronization will be based on the new *scheduled output time* d_n for each stream.

As mentioned earlier, the master stream is often the video stream. Let assume that the scheduled output time of the video stream is the dv_n, and the scheduled output time of the audio, the depth and other sensors stream is da_n.

All the first packets from each stream should get the same output time, which means:

$$dv_1 = da_1 \tag{7}$$

This means that based on Eq. 3, the first ideal output time for the video xv_1 and the other sensors xa_1 should be:

$$xv_1 = xa_1 \tag{8}$$

As the video stream was set as the master stream, its ideal output time should not be changed. This means that if xv_1 is bigger than xa_1, the secondary stream should be delayed and have as final output time xa_1^*

$$xa_1^* = xv_1 \tag{9}$$

If the video stream has smaller ideal output time for its first frame xv_1 than the other streams, the video stream should not be delayed. The other streams should get ideal output time for its first frame

$$xa_1^* = \begin{cases} Aa_1, & \text{if } Aa_1 \geq xv_1 \\ xv_1, & \text{otherwise} \end{cases} \tag{10}$$

where Aa_1 is the arrival time of the first sensor frame.

All the modified output time for the other frames are based on Eq. 4, 5, 6.

5 Wireless Network Infrastructure

All the data from the remote surveilled environment to the data center are transferred wirelessly. All wireless communication standards that could be used for this transmission are shown in Table 3.

Table 3. Wireless Communication Standards

	Max Range (m)	***Max. Upload Data Rate (MBits)***	***Max Power Consumption (mW)***	***Frequency (MHz)***
ZigBee [22]	70 m	0.25	30	2400
Bluetooth [22]	100 m	1	100	2400
802.11ac [23]	150 m	1300	1000	5000
4G [24]	Cellular based	75	200	900/1800/ 2300
5G [25]	Cellular based	1 GBps	Lower than 4G	Undifined

The most promising (although not yet available) wireless standard is 5G (5^{th}-generation mobile network, or 5^{th}-generation wireless system). Stakeholders claim it will be available by 2020. Its target is to provide a bandwidth that is greater than 1 GBps, end-to-end delay smaller than 1 ms, and higher efficiency than its predecessor, 4G.

The prevailed and available standards of Table 3 are the 4G and the 802.11ac because of the available range and data rate that they offer. The 4G uses the cellular network to offer unlimited range. On the other hand, the range of a Wi-Fi network may be extended to only some several hundred of meters but is a lot faster than the 4G. If the distance between the surveilled environment and the access point of an available WiFi is some dozens of meters then the 802.11ac standard is preferred. If the surveilled environment is at a distant location, then the 4G standard is used.

6 A New Transmission Protocol for Multisensory Surveillance Systems

In this paper, a new Network-Adaptive, Multisensory, Real-time, Transmission Protocol (NAMRTP) is proposed. Its main target is to send reliably, real-time multisensory information from the remote surveilled environment to the data center. In order for the protocol to be network adaptive, the delay time and the packet loss are being recorded at the receiver. The protocol changes its sending rate and packet size according to the network conditions. Its primary target is to send all the data reliably without forcing the network to congestion.

6.1 Adaptive Packet Frame Grouping (APFG)

One method that the NAMRTP uses to avoid congestion is Adaptive Packet Frame Grouping (APFG) [26]. When network conditions are deteriorating, the protocol lowers its frame rate transmission in order to avoid congestion. Given that a sensor usually produces data packets at a steady rate, the protocol should group these packets into frames in order to lower the frame rate transmission. The number of grouped packets is being changed according to the network conditions. When network conditions deteriorate, the number of the grouped packets is increased, in order the frame rate to be reduced. As the grouped packets share the same transport and network header, the total application throughput is reduced. The APFG is more preferably from the bandwidth throttling and traffic shaping techniques that are performed by lower-level communication protocols as the APFG technique reduces the application throughput without enforcing packet drops and large packet delays.

The maximum number of grouped packets is depending to the update rate of the sensor, the minimum acceptable refresh rate at the receiver and the maximum acceptable delay that a service can tolerate in order not to affect its real-time feature. Based on the Mean Opinion Score (MOS) on the Quality of Experience (QoE) measurements [27], the Quality of Service (QoS) requirements for real-time services is depicted in Table 4.

Table 4. QoS requirements for real-time multimedia data

QOS	*VIDEO*	*AUDIO*	*GRAPHICS*
JITTER (ms)	≤ 30	≤ 30	≤ 30
DELAY (ms)	≤ 400	≤ 150	≤ 300
PACKET LOSS (%)	≤ 1	≤ 1	≤ 10
REFRESH RATE (Hz)	≥ 30	≥ 50	≥ 30

As most sensor data are illustrated through graphics, the QoS requirements for the NAMRTP protocol are depicted in the third column of Table 4.

If the sensor update rate is *SUR* packets per second, and the maximum affordable delay is 300 ms, the maximum number of grouped packets np_{max} per frame is:

$$np_{max}=0.3*SUR<512 \quad (11)$$

which correspond to the minimum sending rate of 33.33 Hz. The minimum number of grouped packets is set to

$$np_{min}=SUR/100 \quad (12)$$

packets per frame, which correspond to sending rate of 100 Hz, which is a perfect refresh rate for the visual sense.

The number of the grouped packets *np* is changing according to the Eq. 13. The delay of the network d_{net} can easily be extracted if ICMP packets are sent from the sender to the receiver every $T_{ping}=0.5$ sec. The factor d_{max} is the maximum acceptable network delay. For real-time graphics the maximum acceptable delay is set to $d_{max}=300$ *ms.*

$$np_i=\begin{cases} np_0=np_{min} & \\ np_{i-1}-2>np_{min} & ,dnet<dmax/3 \\ np_{i-1}-1>np_{min} & ,dmax/3<dnet<2*dmax/3 \\ np_{i-1} & ,2dmax/3<dnet<dmax \\ np_{i-1}+2<np_{max} & ,dmax<dnet \end{cases} \quad (13)$$

The delay of the network has been divided to four intervals. The first delay interval, $dnet<dmax/3$, correspond to perfect network conditions. The protocol tries to increase the sending rate rapidly. The second interval, $dmax/3<dnet<2*dmax/3$, correspond to good network conditions and the protocol tries to increase the sending rate slowly. The third interval, $2*dmax/3<dnet<dmax$, correspond to acceptable network conditions and the protocol tries to keep the sending rate steady. The last interval $dmax<dnet$ correspond to unacceptable network conditions and the protocol tries to lower the sending rate in order to avoid congestion.

6.2 Adaptive Quantization

Another method that the NAMRTP uses to avoid congestion is the network-adaptive quantization [28]. The quantization levels of the data values are changing according to the network conditions. When the network conditions are good, the quantization levels increase and vice versa. The maximum quantization level is set to $ql_{max}=65536=2^{16}$, which demand 2 bytes per data value. The minimum quantization level is set to $ql_{min}=256=2^8$ which demand 1 byte per data value. An intermediate value of quantization level is set to $ql_{medium}=4096=2^{12}$, which demands 1.5 bytes per data value. The quantization level *ql* is changing according to the Eq. 14.

$$ql=\begin{cases} 65536 & ,dnet<dmax \\ 4096 & ,dnet>dmax\ (For\ t>T_{maxdelay}) \\ 256 & ,dnet>dmax\ \ (For\ t>2*T_{maxdelay}) \end{cases} \quad (14)$$

If the delay of the network is bigger than the maximum acceptable delay d_{max} for a period of time $T_{maxdelay}=2.5$ *sec*, then the quantization level is degraded to 4096. If the delay of the network remain bigger than d_{max} for a period of time greater to 2* $T_{maxdelay}$, then the quantization level is degraded to 256.

6.3 NAMRTP Packet Header

The NAMRTP protocol runs over the UDP protocol. The UDP protocol is chosen as it is lighter and faster than the RTP and the TCP transport protocol. The progressive http transmission was rejected as it introduces unwanted and unnecessary traffic and delay to the network due to TCP acknowledgements and buffering. The NAMRTP is used to compensate the UDP's lack of reliability and congestion control. The NAMRTP is lighter and more efficient than the multipurpose, widely-used for real-time services RTP protocol, as it is focused only in sensory data transmission. The header of the NAMRTP protocol is by 4 bytes smaller than the RTP's header and is illustrated in Figure 3.

Bits	***0 - 15***	***16-18***	***19-23***	***24-31***
0	Sequence Number	QuanL.	SenId	NumPack.
32	Timestamp			
64	Data			

Fig. 3. Header of an NAMRTP protocol packet.

The Sequence Number is being used to reinforce the reliability of the protocol through the acknowledgment process.

The QuanL. (Quantization Level) field is informing the receiver for the quantization level of the data and the length of each data value.

The SenId (Sensor ID) field informs the receiver to which sensor the data correspond to.

The NumPack (Number of Packets) field informs the receiver the number of packets that are grouped in this frame. As the NumPack is consisted by 8 bits, the maximum number of grouped packets is 512.

The Timestamp field is needed for intra/inter synchronization control.

6.4 NAMRTP Reliability

NAMRTP protocol apart from being efficient and network adaptive, it should also be reliable. As a real-time protocol, it should timely deliver the packets. But as the packets include surveillance information that could be used for post processing, all the packets should be transferred reliable.

In order to enforce reliability, a sequence number is used. Every frame has a unique sequence number. The method of the cumulative negative acknowledgement (CNAK) [29, 30] is applied. The sender sends the frames as soon as they are created. A copy of these frames is stored in a buffer at the sender side in case a negative acknowledgement is received. A CNAK is sent by the receiver every *k= 1 sec* containing the frames that have not been received. The receiver also sends the last sequence number that has received to help the sender to empty its buffer.

The packet construction for the cumulative negative acknowledgment is depicted in Figure 4.

Bits	***0-3***	***4-7***	***8 - 15***	***16-31***
0	Sensor Id	ACK Sequence Number		Last Received Sequence Number
32	Data (Dropped Sequence Numbers)			

Fig. 4. Header of an NAMRTP CNAK packet.

This CNAK packet should be sent reliable. An acknowledgment over the UDP protocol, containing the lost frames, is sent by the sender as soon as it receives a CNAK. If the receiver doesn't get this acknowledgment in a specific time period *h= 1 sec*, a CNAK is resent.

The flow diagram of the NAMRTP protocol is depicted in Figure 5.

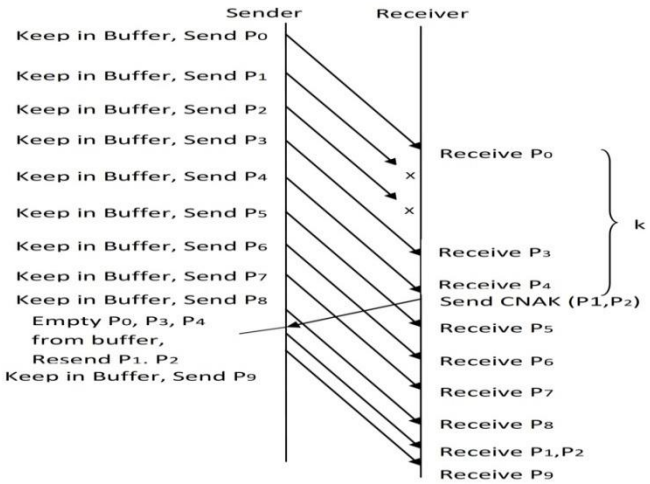

Fig. 5. Flow diagram of the NAMRTP protocol.

7 Measurements for Real-Time Transmission of 3D-HEVC Multisensory Streams

The NAMRTP protocol was tested in two different network topologies, with two different sensors and two different network traffic scenarios. In all cases, the performance of the NAMRTP protocol was compared against the fast but unreliable UDP without Adaptive Packet Frame Grouping (APFG) and adaptive quantization. The performance metrics that were measured were the packet loss, the average end to end delay, and the standard variation of the delay, also known as jitter.

In the first topology, the surveillance sensor was sending its data through an 802.11n WiFi 300-Mbps wireless network. In the second topology, the surveillance sensor was sending its data through a 3G HSDPA network.

Two different sensors were used. The first sensor had an update rate of 30 packets-per-second, which is usually encountered in depth cameras. The second sensor was

producing packets at a rather high update rate of 1000 Hz. The value size of each sensor packet was changing based on Eq. 14 from 1 to 2 bytes according to network conditions. The packet frame grouping was changing based on Eq. 11-13.

Two different network traffic scenarios were used. In the first case, no other network traffic were sent over the wireless/ 3G network. In the second case an HEVC 1080p at 30 fps video and an AAC-LC with 96-Khz sample rate audio stream was sent simultaneously with the sensor stream over the RTSP protocol.

Three measurements for each case of the above transport protocols, network topologies, sensors and network traffic scenarios have been taken. For each measurement 10000 frames where sent. The best results from each of the above measurements are depicted in Table 5 and 6.

Table 5. NAMRTP and UDP transmission over WiFi and 3G network with no other network traffic

TOPOLOGY	*PROTOCOL*	*SENSOR UPDATE RATE Packet/sec*	*AVG DELAY ms*	*JITTER ms*	*PACKET LOSS %*
WiFi	AMRTP	1000	2.36	1.355	0.15
WiFi	AMRTP	30	4.45	10.58	0
WiFi	UDP	1000	3.74	22.89	0.27
WiFi	UDP	30	4.49	12.74	0.1
3G	AMRTP	1000	70.51	97.64	0.15
3G	AMRTP	30	60.29	88.1	0.99
3G	UDP	1000	248.96	261.58	29.74
3G	UDP	30	57.16	35.83	0

Table 6. NAMRTP and UDP transmission over WiFi and 3G network simultaneously with a 1080p video stream

TOPOLOGY	*PROTOCOL*	*SENSOR UPDATE RATE Packet/sec*	*AVG DELAY ms*	*JITTER ms*	*PACKET LOSS %*
WiFi	AMRTP	1000	2.88	3.3	0
WiFi	AMRTP	30	3.19	3.58	0
WiFi	UDP	1000	2.85	2.16	0
WiFi	UDP	30	2.50	2.7	0
3G	AMRTP	1000	172.31	323.29	0
3G	AMRTP	30	114.02	213.60	0
3G	UDP	1000	570.67	402.39	26.66
3G	UDP	30	202.89	332.11	0

It is understood that in both network traffic scenarios WiFi network conditions are by far better than the 3G network. The average delay and jitter are smaller to the WiFi network.

Average end to end delay is almost invariant to the enforced transport protocol, to the sensor update rate and the traffic network scenario in the case of WiFi network topology.

In the case of the 3G network, the average delay increases when the sensor update rate is increased.

The packet loss in most circumstances is below 1%. The only scenario that shows increased packet loss is the UDP transport protocol over the 3G network with sensor update rate equal to 1000 packets per sec. The packet loss rate in these circumstances is 26-29%, which is unacceptably high. This high packet loss is can be lowered if the sending rate is reduced. When the UDP transport protocol is used, and the sending rate changes from 1000 packets per second to 30, Table 5 and 6 depicts that the packet loss is reduced from 26.66 and 29.74% to 0%. This means that the 3G network is not capable to transfer 1000 packets per second, and many packets are dropped. This observation is being exploited by the NAMRTP protocol and reduces the sending rate by grouping the packets to frames. That's why the packet loss in the case of the NAMRTP is so small.

As the packet loss is almost zero in the case of the NAMRTP protocol, the cumulative negative acknowledgment is the best reliability mechanism, as it reduces network traffic from the unnecessary acknowledgments. All the packets that are lost in the NAMRTP protocol are resend successfully with the reliability mechanism.

Comparing the NAMRTP to the UDP transport protocol, it is understood that they have almost the same results in the case of the WiFi network. The advantage of the NAMRTP protocol is its reliability, due to the cumulative negative acknowledgement that it enforces.

In the 3G network, the NAMRTP protocol outperforms the UDP protocol in almost all the performance metrics. As the network conditions of the 3G network are inadequate for the transmission of high update rates, the adaptive packet frame grouping and the adaptive quantization of the NAMRTP protocol enhance the performance of the transport protocol.

Apart from the main performance metrics, delay, jitter and packet loss, the adaptive behavior of the frame grouping and the quantization was monitored.

In the case of the WiFi network topology, the network conditions were rather adequate for the multisensory streams. The number of the grouped packets per frame was constantly equal to the minimum number of grouped packets np_{min} based on Eq. 12. Similarly, the quantization level of Eq. 14 remained equal to its maximum level in all WiFi experiments.

On the other hand, the 3G network with video traffic was rather congested. The number of the grouped packets and the packet size where fluctuating according to the network conditions, in order to avoid congestion.

Figure 6 depicts the adaptive number of grouped packet per frame for the 3G network with and without network traffic.

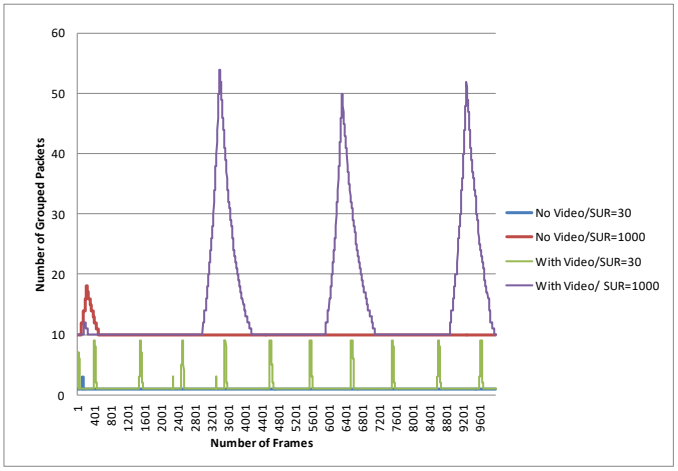

Fig. 6. Adaptive number of grouped packets per frame.

Based on Eq. 12, when the Sensor Update Rate (SUR) is equal to 30, the minimum number of grouped packets is 1. When the SUR is equal to 1000 the minimum number of grouped packets is 10. It is obvious that when video traffic is enforced, the network is rather congested and the number of grouped packets often increases.

Figure 7 depicts the adaptive quantization level for the WiFi and the 3G network with and without network traffic.

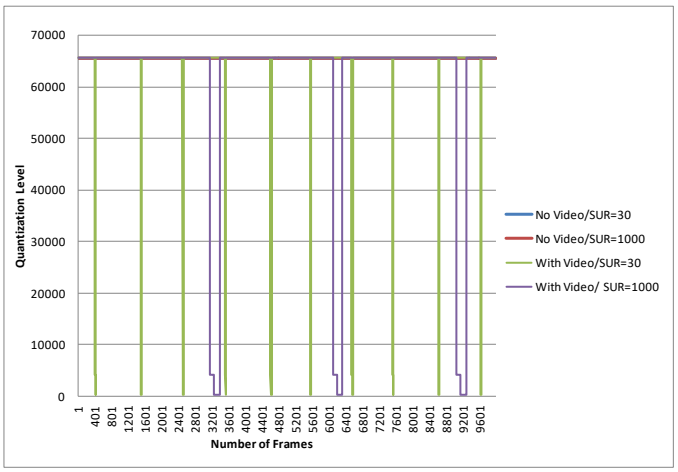

Fig. 7. Adaptive quantization level.

When the 3G Network doesn't transfer any video traffic, the quantization level is constant at its maximum value 65536. When the 3G network transfers video traffic the quantization level fluctuates from 65536 to 256, in order to lower the packet size, and avoid congestion and packet loss.

8 Conclusions

This paper describes the architecture of a real-time, wireless, multisensory smart surveillance system. All possible data streams are presented. A data center is proposed as both a data storage repository as well as a data processing and prediction unit. Future threats are predicted and real-time alerts are generated.

The advantages of HEVC over its predecessor, H.264, for video stream transmission and storage in surveillance systems are presented. 3D monitoring techniques are proposed. Intra- and inter-frame synchronization techniques for real-time multisensory streams are analyzed and deployed.

Wireless communications standards are investigated and the advantages of 802.11ac/n and 4G networks are analyzed.

A novel, reliable, network-adaptive, transmission protocol, named NAMRTP, is proposed. It successfully adapts the transmission rate and bandwidth by using Adaptive Packet Frame Grouping (APFG) and adaptive quantization in order to reduce congestion and packet loss.

Experimental tests have shown that the proposed protocol is a promising candidate for real-time, multisensory surveillance systems over Wi-Fi or 4G networks. The reliability and congestion control mechanisms used by the NAMRTP protocol allow for reliable data transfer over time-varying network conditions.

References

[1] I. D. N. Enterprise, R. I. G. Micro, Nanosystems, and W. G. R. of the ETP EPOSS, "Internet of things in 2020: Roadmap for the future," Version 1.1, Tech. Rep., 27 May 2008.

[2] R. Garcia and H. Kalva, "Subjective evaluation of hevc and avc/h.264 in mobile environments," *IEEE Tran. Consumer Electronics*, vol. 60, no. 1, pp. 116–123, February 2014.

[3] Valera, Maria, and Sergio A. Velastin. "Intelligent distributed surveillance systems: a review." IEE Proceedings - Vision, Image and Signal Processing, Vol. 152. No. 2. IET, 2005.

[4] Kim, Hyung-Il, Seung Ho Lee, and Yong Man Ro. "Low-Power Face Detection for Smart Camera", *Theory and Applications of Smart Cameras*, 2016. pp 139-155.

[5] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A system for video surveillance and monitoring," Carnegie Mellon University Technical Report, CMU-RI-TR- 00-12, 2000.

[6] N. T. Siebel and S. Maybank, "The advisor visual surveillance system." Proc. of the ECCV Workshop on Applications of Computer Vision, pp. 103-111, 2004.

[7] C. F. Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior, and Y. Tian, "IBM smart surveillance system (S3): an open and extensible framework for event based surveillance," Proc. Of IEEE Conf. Advanced Video and Signal Based Surveillance, pp. 318-323, 2005.

[8] C. Regazzoni, V. Ramesh, and G. L. Foresti, "Special issue on video communications, processing, and understanding for third generation surveillance systems," *Proc. of the IEEE*, vol. 89, no. 10, pp. 1355-1367, 2001.

[9] X. Dong, J. Wen, "Low lighting image enhancement using local maximum color value prior", *Frontiers of Computer Science*, vol 10 no.1, pp 147-156, 2016.

[10] S. Sanjay, C. Shekhar, and A. Vohra. "FPGA-Based Real-Time Motion Detection for Automated Video Surveillance Systems", *Electronics* , vol. 5, no. 1, pp. 10, 2016.

[11] D. J. Guo, L. Zhe-Ming, and L. Hao, "Multi-Channel Adaptive Mixture Background Model for Real-time Tracking", *Journal of Information Hiding and Multimedia Signal Processing*, vol. 7, no. 1 , pp 216-221, 2016.

[12] A. E. Maadi, and M. S. Djouadi, "Large-scale surveillance system: detection and tracking of suspicious motion patterns in crowded traffic

scenes", *Automatika–Journal for Control, Measurement, Electronics, Computing and Communications*, vol 60. no. 1, 2016.

[13] S. Zhengguo, Y. Shusen; Yifan Yu, A. Vasilakos, J. McCann, L. Kin, "A survey on the IETF protocol suite for the internet of things: standards, challenges, and opportunities", *IEEE Wireless Communications*, vol.20, no.6, pp.91-98, December 2013

[14] J. Mongay Batalla, "Advanced multimedia service provisioning based on efficient interoperability of adaptive streaming protocol and high efficient video coding," *Journal of Real-Time Image Processing*, pp. 1–12, 2015.

[15] E. Kurniawati, C. Lau, B. Premkumar, J. Absar, and S. George, "New implementation techniques of an efficient mpeg advanced audio coder," *IEEE Tran. on Consumer Electronics*, vol. 50, no. 2, pp. 655–665, May 2004.

[16] J. Ohm, G. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards;including high efficiency video coding (hevc)," *IEEE Tran. Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, Dec 2012.

[17] F. Fitzek and M. Reisslein. Video traces for network performance evaluation: Yuv 4:2:0 video sequences. [Online]. Available: http://-trace.eas.asu.edu/yuv/yuv.html

[18] G. Kokkonis, K. E. Psannis, M. Roumeliotis, and Y. Ishibashi, "Efficient algorithm for transferring a real-time hevc stream with haptic data through the internet," *Journal of Real-Time Image Processing*, pp. 1–13, 2015.

[19] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb 2012.

[20] P. K. Nathan Silberman, D. Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[21] Y. Ishibashi, T. Kanbara, and S. Tasaka, "Inter-stream synchronization between haptic media and voice in collaborative virtual environments," in *Proc. of the 12th Annual ACM Int. Conf. on Multimedia*, 2004, pp. 604–611.

[22] J.-S. Lee, Y.-W. Su, and C.-C. Shen, "A comparative study of wireless protocols: Bluetooth, uwb, zigbee, and wi-fi," in *Industrial Electronics Society, 2007. IECON 2007. 33rd Annual Conference of the IEEE*, Nov 2007, pp. 46–51.

[23] E. H. Ong, J. Kneckt, O. Alanen, Z. Chang, T. Huovinen, and T. Nihtila, "Ieee 802.11ac: Enhancements for very high throughput wlans," in *IEEE 22nd Int. Symp. on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept 2011, pp. 849–853.

[24] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4g lte networks," in *Proc. 10th International Conference on Mobile Systems, Applications, and Services*, 2012, pp. 225–238.

[25] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[26] J. Tourrilhes, "Packet frame grouping: improving ip multimedia performance over csma/ca," in *IEEE Int. Conf. on Universal Personal Communications, 1998*, vol. 2, Oct 1998, pp. 1345–1349.

[27] K. Iwata, Y. Ishibashi, N. Fukushima, and S. Sugawara, "Qoe assessment in haptic media, sound, and video transmission: Effect of playout buffering control," *Computers in Entertainment (CIE)*, vol. 8, no. 2, p. 12, 2010.

[28] L. Guiyun, J. Yao, Y. Liu, H. Chen, and D. Tang, "Channel-aware adaptive quantization method for source localization in wireless sensor networks," *International Journal of Distributed Sensor Networks*, no. 214081, p. 13, 2015.

[29] J. Baek and C. Kim, "An energy-efficient video transport protocol for personal cloud-based computing," *Journal of Real-Time Image Processing*, pp. 1–8, 2014.

[30] E. Mulabegovic, D. Schonfeld, and R. Ansari, "Lightweight Streaming Protocol (LSP)", Proc. 10th ACM Int. Conf. on Multimedia, MULTIMEDIA '02, pp. 227-230, 2002.