Cluster Computing (2022) 25:231–247
https://doi.org/10.1007/s10586-021-03391-4

Towards the availability of video communication in artificial intelligence-based computer vision systems utilizing a multi-objective function

Yousef O. Sharrab[1,2] · Izzat Alsmadi[3] · Nabil J. Sarhan[4]

Received: 3 February 2021 / Revised: 3 August 2021 / Accepted: 7 August 2021 / Published online: 26 August 2021

Abstract
Availability is one of the three main goals of information security. This paper contributes to systems' availability by introducing an optimization model for the adaptation (controlling the capturing, coding, and sending features of the video communication system) of live broadcasting of video to limited and varied network bandwidth and/or limited power sources such as wireless and mobile network cases. We first, analyzed the bitrate-accuracy and bitrate-power characteristics of various video transmission techniques for adapting video communication in Artificial Intelligence-based Systems. To optimize resources for live video streaming, we analyze various video parameter settings for adapting the stream to available resources. We consider the object detection accuracy, the bandwidth, and power consumption requirement. The results showed that setting SNR and spatial video encoding features (with upscaling the frames at the destination) are the best techniques that maximizing the object detection accuracy while minimizing the bandwidth and the consumed energy requirements. In addition, we analyze the effectiveness of combining SNR and spatial video encoding features with upscaling and find that we can increase the performance of the streaming system by combining these two techniques. We presented a multi-objective function for determining the parameter or parameters' pairing that provides the optimal object detection's accuracy, power consumption, and bit rate. Results are reported based on more than 15,000 experiments utilizing standard datasets for short video segments and a collected dataset of 300 videos from YouTube. We evaluated results based on the detection index, false-positive index, power consumption, and bandwidth requirements metrics. For a single adaptive parameter, the analysis of the experiment's outcome demonstrate that the multi-objective function achieves object detection accuracy as high as the best while drastically reducing bandwidth requirements and energy consumption. For multiple adaptive parameters, the analysis of the experiment's outcome demonstrate the significant benefits of effective pairings (pairs) of adaptive parameters. For example, by combining the signal-to-noise ratio (SNR) with the spatial feature in H.264, a certain optimal parameter setting can be reached where the power consumption can be reduced to 20%, and the bandwidth requirements to 2% from the original, while keeping the Object Detection Accuracy (*ODA*) within 10% less of the highest *ODA*.

Keywords Video communications · Optimization · Computer vision systems · Power consumption · Rate-energy-accuracy trade-offs · Video stream adaptation · Artificial intelligence-based systems · Availability of video communication · Multi-objective function · AI-based computer vision systems

1 Introduction

Computer vision (CV) is the application of Artificial Intelligence (AI) to the visual world [1]. It describes the ability of machines to process and understand visual data like images and videos. This paper deals with the design of real-time AI-based computer vision systems (*AICV*) in which objects and events are analyzed automatically by running computer vision algorithms. Availability is one of the three main pillars of information security [2]. The paper contributes to the availability of *AICV* systems by presenting an optimization model for adapting live video communication to the limited and varied network resources of bandwidth and power. One example of such systems is

Extended author information available on the last page of the article

large-scale Intelligent video surveillance(*IVS*) [3]. In such systems, several video sources (cameras) broadcast the video clips to a central monitoring station for analysis by running CV algorithms. In addition, Human-to-human video communication applications such as video conferencing, distance learning, and telecollaboration can utilize our multi-objective model for optimal bandwidth usage and/or energy consumption. In these applications, the accuracy can be mapped to human perceptual quality metrics. Due to its nature, the main performance metric in the *AICV* systems is the object detection accuracy of the CV algorithm(s) used. Energy consumption has become a main worry [4]. Hence, it should be considered as well as bandwidth. *AICV* systems have limitations, including those in power and bandwidth. Therefore, Several video streams must be dynamically adapted based on available resources and the maximum object detection accuracy. Adapting video is achieved by changing the capture and encoding parameters of Several video sources. These parameters include resolution, frame-rate, and bit rate such as Quantization Parameter (QP) or Scaling Factor (SF).

Few research considered AI-based computer vision systems. Study [5] considered the impact of adapting videos on the detection accuracy of faces but discussed only MJPEG videos and utilized a small dataset. In addition, the paper did not include energy consumption. Studies [6] considered bandwidth allocation in video surveillance systems but did not analyze Several video encoding features.

This paper studies Several video communication techniques for adapting parameters according to accuracy of object detection, bit rate (bandwidth), and energy consumption. Adapting video streams means adapting the capture, encoding, and transmission parameters of the video stream. Parameters include Signal-to-Noise Ratio (*SNR*) and spatio-temporal resolution. The analyzed techniques for adapting videos include *spatial* (resolution), *temporal* (frame rate), and *SNR* (quantization parameters/scaling factor for H.264, MPEG-4, respectively) as discussed in Sect. 2. In addition, the paper discusses the effect of the receiver station upscaling (frame-enlargement) for the frames of videos that are spatially adapted to their original frame size (resolutions) utilizing super-resolution upscaling techniques [7, 8] and then applying the computer vision algorithm. In addition, the paper discusses the effect of several pairs of encoding parameters.

In addition, this paper presents a multi-objective function and algorithm to select an adaptive or adaptive set that provides the best overall trade-off according to computer vision object detection accuracy, power consumption, and bit rate. Furthermore, the multi-objective function checks the rate of change in each, thus giving preference to the setting with the larger subsequent drop in resolution, the smaller subsequent decrease in bit rate, and/or the smaller subsequent decrease in power.

We report results based on more than 15, 000 real experiments, on a dataset of 300 videos collected from YouTube and additional standard video segments . The total recording time for these videos is over 19. We experiment on two types of systems. The first includes an FFmpeg-encoded computer with an external camera, while the other includes a real surveillance camera with encoding implemented by System-on-Chip (SOC). We analyze the recordings in both MPEG-4 and H.264 encoders and assess several metrics.

The *main contributions* of this paper can be summarized as the followings. (1) To optimize resources for live video streaming toward the availability of the system, we analyze various video parameter settings for adapting the stream to the available resources. Considering the object detection accuracy, bandwidth requirements, and power consumption, we find that setting SNR and spatial video encoding features with upscaling the frames at the destination are the best techniques that maximize the object detection accuracy, minimize the bandwidth and the consumed energy requirements. (2) We analyze the effectiveness of combining SNR and spatial video encoding features when the destination station employs upscaling (enlarging) and find that we can increase the performance of the streaming system by combining these two techniques much more than utilize them individually. The better performance is the less bandwidth and energy requirements for the maximum object detection accuracy. We decrease the bandwidth and energy requirements for video streaming, while keeping the object detection accuracy at the destination without significant degradation. (3) We develop a multi-objective function and an optimization strategy for determining the optimal system feature technique or features pairing.

For a single adaptive technique, the analysis of the experiment's outcome show that the multi-objective function achieves objective accuracy close to the best, while significantly reducing bandwidth requirements and consumed energy. For multiple video encoding features techniques, the analysis of the experiment's outcome demonstrate the amount of benefits of combining of effective features techniques. As an example, by combining the signal-to-noise ratio (SNR) with the spatial video encoding features in H.264, optimal setup can reduce the power consumption to 20%, and the bandwidth requirements to 2% from the original, while keeping the Object Detection Accuracy (*ODA*) within 10% from maximum.

A small portion of this paper is based on our preliminary work in a conference paper [9].The paper studies the resolution and power versus bit rate characteristics of Several video video encoding features in computer vision systems.

The analyzed video encoding features include spatial (and spatial with upscaling), temporal, and *SNR*.

The analysis of the experiment's outcome demonstrate that SNR system feature generally resulted in the best characteristics of the rate-accuracy, followed by spatial with upscaling. Therefore, we in this paper experiments only with Signal-to-Noise Ratio (*SNR*) and spatial (with upscaling at receiver) video encoding features, we analyzes how video encoding features can be combined, presents the framework for selecting video encoding features, proposes the multi-objective function, and provides comprehensive evaluations using Several experimental setups. Only the first three figures in Sect. 5 are based on that work and the rest are completely new. The main symbols used in the paper are summarized in Table 1.

The remaining parts of the paper is organized as following: Background information and related work is discussed in Sect. 2. Section 3 discusses the proposed algorithm. The methodology is discussed in Sect. 4, presenting and analyzing the analysis of the experiment's outcome are in Sect. 5. Finally, in the last section the conclusions are drawn.

Table 1 Summary of notations

Notation	Descriptions
ABE_MOF/ ABE_OF	Accuracy-bitrate-energy multi-objective/objective function
m	Weight for normalized accuracy
p	Weight for normalized bit rate
u	Weight for normalized consumed energy
n	Weight for normalized accuracy rate-of-change
q	Weight for normalized bit rate rate-of-change
v	Weight for normalized consumed energy rate-of-change
A	The normalized detection accuracy
B	The normalized detection bit rate
E	The normalized consumed energy
g_A	The normalized detection accuracy rate-of-change
g_B	The normalized bit rate rate-of-change
g_E	The normalized consumed energy rate-of-change
f(x, y)	Rate-of-change for the cell x,y in the video encoding features matrix
T_c	Time complexity
O()	Big-O complexity
N_adaptations	Number of video encoding features
N_frames	Number of video frames
R_x	Resolution in the x dimension
R_y	Resolution in the y dimension

2 Background information and related work

Most of the research in AI-based computer vision systems has focused on developing powerful computer vision algorithms to detect and track objects [10, 11] and the object detection and tracking of unusual events [12]. However, a little work considers the availability of video communication in such systems.

Sending video over the network requires a large amount of bandwidth because video files usually are very big. The video streams might need to adapte to available resources [13]. To reduce the file size so that it requires less transmission bandwidth, the video is encoded/compressed. The file size after encoding is determined by the size or the bandwidth requirements of the encoded video. Bitrate means how much bits are required to transmit at each second. There are several video encoding standards such as H.264, MJPEG, MPEG-4, VP8, VP9, and HEVC.

The basic methods of video stream adaptation include scaling of video signal-to-noise ratio (*SNR*), temporal and/or spatial parameters. The *SNR* quality practically is the intensity of pixels, which is controlled by a certain parameter for each encoder, like Scaling Factor (SF) in JPEG and Quantization parameter in MPEG-4. However, temporal (the rate of frames) and spatial (the size of the frame) qualities are controlled by changing the rate of frames (the total number of frames per second) and the the size of the frame (resolution or the total number of pixels in each frame), respectively. The frame size can be shrinked by dropping pixels from the frames of the the video while the rate of frames can be decreased by dropping frames from the video. At the receiver side, the video is decoded. In *AICV* systems, adaptation involves choosing the capture, encoding, and transmission parameters required for the camera.

In video encoding, the frames are grouped in sequences. The first frame is encoded utilizing spatial similarities (intra-coding). In intra-coding, as in the first frame of a sequence, the prediction of each block is utilized by neighboring blocks that are previously encoded. In a sequence, the other remaining frames are encoded based on the temporal similarities in an encoding process called inter-coding. In inter-coding, the prediction of the blocks is based on motion estimation and compensation of the block compared to the corresponding block in the previously encoded frames. Then it is the residuals(the difference between the predicted and the original blocks) of the intra-coding and inter-coding transformed to the frequency domain. Then, to eliminate high-frequency coefficients and to reduce the overall precision of the coefficients, the transform coefficients are quantized. Subsequently, these

coefficients are compressed utilizing entropy coding and then transmitted in addition to motion vectors (MVs).

The encoders are developed to reduce the required bandwidth and to keep the quality for human eyes. In Artificial Intelligent and computer vision systems we care about the detection accuracy instead of quality for human eyes [14].

For spatially-adapted videos (i.e., videos whose spatial resolution or number of pixels in the frames is reduced), To enlarge the video frames at the receiver, an algorithm can be used to upscale them to obtain higher resolution through computer vision algorithms. Finally, computer vision algorithms will be run on the incoming video streams.

In upscaling (also called super-resolution), interpolation is used to improve the resolution of a video frmaes or images. We analyze the effectiveness of the most popular upscaling algorithms in this paper, which include: *Lanczos*, *Spline*, *Bicubic*, *Bilinear*, *Nearest Neighbor*. The last three algorithms consider the closest 4×4 pixels, the closest 2×2 pixels, and the closest pixel, respectively. Lanczos and Spline algorithms take into account additional surrounding pixels. Other algorithms are in [15, 16].

Live video communication and streaming systems consist of three main stages at the source side: capture, compression/encoding, and transmission. We discuss each of these stages in the next subsections.

Computer vision (*CV*) is an area of artificial intelligence that trains computers to perceive and understand the visual world. Using digital images from cameras, videos, and deep learning models, machines can accurately detect, recognize, and track an object [17]. Major *CV* functions are object detection, recognition, and tracking [18]. Computer vision systems and applications are developed based on these functions, and the required software, hardware,and computer networks.

Artificial Intelligence-based Computer Vision System have limitations in computational, communication, and power consumption . A measure (metric) of the performance of computer vision systems is their functions accuracy. Video encoding power optimization is one of the most challenging design factors in video communication. The coder parameters settings used at the transmitting devices has high impact on bandwidth requirement, consumed energy, and video clarity.

Intelligent video security monitor alert the operator of unwanted activities. The Computation is extremely high due to the processing of video frames. Encoding in these video systems must be done in real time. The video clarity, bandwidth requirements, and compression time are determined by several encoding parameters. These parameters must be controlled to adapt based on the available bandwidth and energy for efficient encoding. [19, 20].

Most research on artificial intelligence-based computer vision (*AICV*) has focused on developing powerful computer vision algorithms for classification, object detection, and tracking [21–27]. However, less work has focused on system design, and specifically, no research has been considered the bandwidth, power consumption, function accuracy optimization in such systems.

As Fig. 1 illustrate, the considered Artificial Intelligence-based Computer Vision System (*AICV* systems) includes several cameras, which transmit videos to a

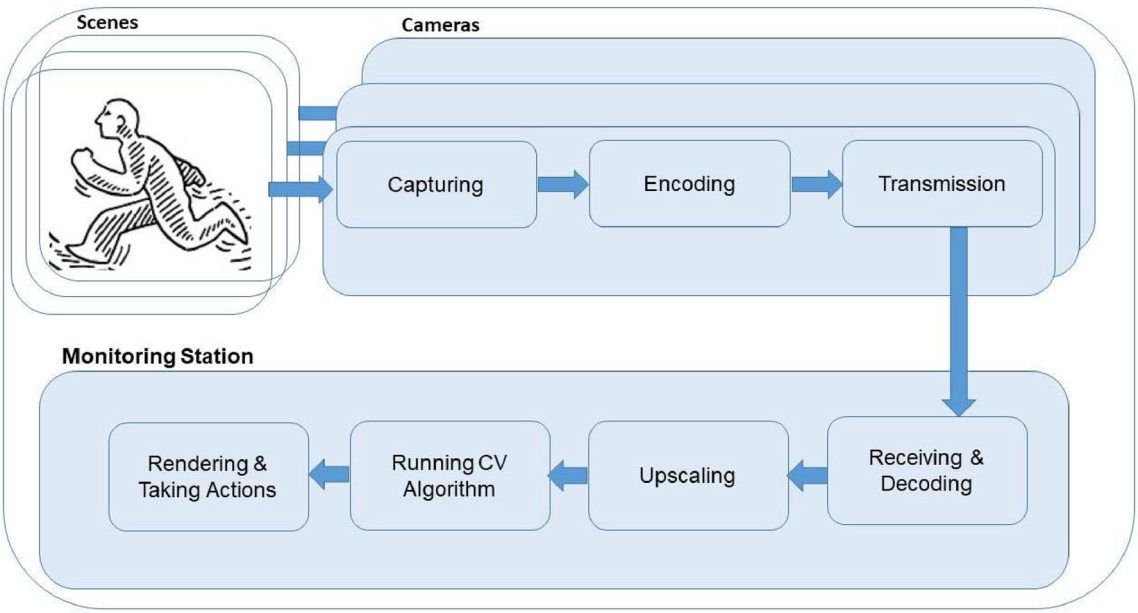


Fig. 1 An illustration of the considered artificial intelligence-based computer vision system)

receiving station. One of the main applications of this system is automated video surveillance (*AVS*). To enable live video streaming, each source goes through three main stages: capturing, compression (encoding), and sending. Because of the complexities of estimation prediction, the encoding stage has the highest CPU computation and power consumption [28, 29]. At the receiver, the video streams are decoded and displayed.

3 Adapting the video stream to the available resources

To adapt the video communication system, we select the capturing, video compression parameters for the capturing device (camera) to suit bandwidth and/or power consumption constraints while maximizing the CV function accuracy. How to transform the video to meet the transmission resource constraint is the function of *adapting video streams* [13].

We analyze the main methods of adapting video stream, including changing the video pixels quality (SNR), resolution (frame size), and/or frame rate. Changing the intensity of pixels (*SNR*) is controlled by changing the scaling factor in MPEG-4 codec or by the quantization parameter (*QP*) in H.264 codec standard. However, video qualities that produced by temporal and spatial adaptation are controlled by frame rate and resolution, respectively.

Videos that have their frame sizes reduced at the sending station can be up-scaled to larger clearer sizes by receivers before the detection/recognition algorithms are applied. In this work, we study and compare the performance of major upscaling algorithms which include: *Bicubic*, *Bilinear*, *Nearest Neighbor*, *Lanczos*, and *Spline*.

3.1 Techniques for adapting the video stream

To develop an efficient *AICV* solutions, we need to analyze several video compression techniques to adapt the the process of compression to the available resources. Our research compares and contrast several tools for adaptive encoding for function accuracy, energy consumed by the encoder, and the required bandwidth. For computer vision functions accuracy (face detection accuracy), we utilize several metrics: The first, *the average accuracy of detection*, which is the number of detected faces correctly (True-positive) divided by number of faces in the video sequence. The second, *number of face detection in the video segment*, this is un-normalized detection accuracy, we use it for YouTube collected videos since it is hard to count the total number of faces in that kind of videos where there are multiple faces in several frames. The third, *false-positive index*, which shows the possibility of a false positive.

3.2 A proposed multi-objective function for optimal AI-based video communication

We find that the pairing of several video encoding features increases adaptive performance [9]. To optimize the benefit of the parameters pairing, we develop a multi-objective function [30], called *Accuracy-Bitrate-Energy Multi-Objective Function* (ABE_{MOF}), Which is a tool to identify the particular optimal parameter or multiple group of parameters that can be used. Figure 2 shows the process of applying the multi-objective function. This function takes into account the computer vision function accuracy, bit rate, the consumed power, and the partial derivative of each. The input features of ABE_{MOF} is an $N \times M$ *video encoding features matrix*, where rows are several *QPs* and the columns are several frame size. Each entry is a set with CV function accuracy, required bandwidth (bit rate), and predicted consumed energy for the corresponding parameters or set of parameters.

The model ABE_{MOF} input are the video encoding features matrix (adaptation matrix), it applies the weights to produce the results in a *Multi-Objective Matrix*, which includes the all the *Multi-Objective Matrix* value for each video encoding features group, with larger values being the better for system performance. In addition to giving advantage to higher accuracy metric, lower bandwidth requirements, and lower power consumption, the multi-objective function checks the partial derivatives for each of these measures and favors the setting with a next larger decrease in accuracy, a next decrease in bandwidth requirements, and/or a next smaller decrease in power.

We need to maximize the accuracy, minimize the bandwidth requirements, and the consumed energy. Therefore, we build the multi-objective function to be directly proportional to the accuracy, inversely proportional to the bit rate, and consumed energy. Since we prefer the settings that decrease the bit rate or consumed energy without affecting the accuracy, the model will have a higher value, if the accuracy is going to decrease for the next settings. The reason is that we do not want to decrease the setting in order to keep the accuracy high. The opposite is to the change of bit rate and consumed energy, we prefer the setting that decreases them as long as it keeps the accuracy without significant change. Based on this, intuitively the multi-objective function can be formulated as follows:

$$ABE_{MOF} = \frac{(A+1)^m \cdot (|g_A|+1)^n}{(R+1)^p \cdot (|g_R|+1)^q \cdot (E+1)^u \cdot (|g_E|+1)^v}, \quad (1)$$

where the features A, R, E, g_A, g_R, and g_E are the normal computer vision function detection accuracy, normal bit

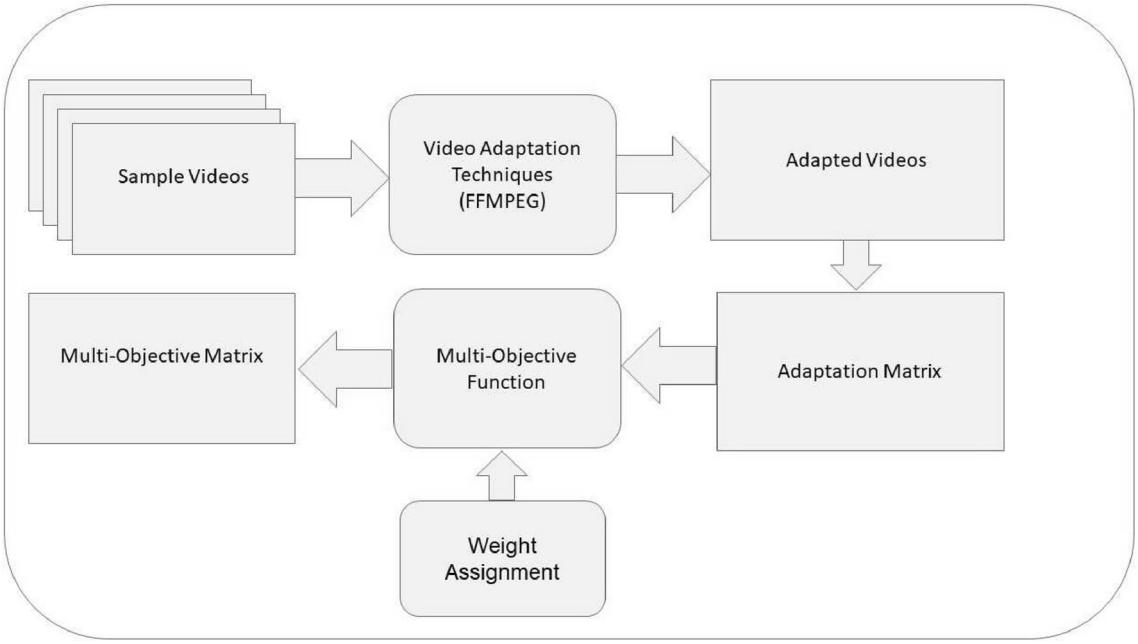


Fig. 2 An illustration of the process of applying the multi-objective function

rate, normal power consumption, and the derivative of each, respectively. Thus, the variables have values in the closed real numbers period [0, 1] (i.e. $\{parameter \mid 0 \leq parameter \leq 1\}$). Constants m, p, u, n, q, and v are assigned weights, with values between and within 0 and 1, the closed real number interval of [0, 1], for the normalized function accuracy, bit rate, and power consumption, and rates of change of each, respectively. Each of these weights is used as an exponent in the model to get a value of 1 (i.e. no impact) for the corresponding feature (accuracy, bit rate, energy, and their derivatives) when this feature is not taken into account by the multi-objective function. We have the option to ignore the derivative of any or all the features (i.e. g_A, g_R, and g_E) in the model (multi-objective function) by assigning zero to n, q, and/or v, respectively. In addition, we can ignore the impact of bit rate and consumed energy by setting p and u to zeros. The value of each term in $(A+1)^m$, $(|g_A|+1)^n$, $(R+1)^p$, $(|g_R|+1)^q$, $(E+1)^u$, and $(|g_E|+1)^v$ is within the closed real interval [1, 2]. The added one to each of these terms is to have no impact on ABE_{MOF} if the associated feature is set to zero in the video encoding features matrix.

To illustrate the impact of the derivatives of the accuracy consideration, let us consider the case when two consecutive rows in the video encoding features matrix are as shown in Table 2. The second entry in the multi-objective matrix will have the largest ABE_{MOF} value (i.e. 0.96) because of the big drop in the corresponding accuracy (matching cell in Accuracy).

The (derivative) diagonal difference of a two-dimensional function $H(A, R)$ of the function H between two successive points $((A, R))$ and $((A+1, R+1))$ can be expressed as follows:

Table 2 An Example Illustrating the Multi-Objective Function [bit rate and consumed energy in kbps and W, respectively. Weights: $m = 1.0, n = 0.1, p = 0, q = 0, u = 0$, and $v = 0$]

Row number	Specific video encoding features dimension	Entry 1	Entry 2	Entry 3	Entry 4
Video encoding features matrix row 1	Accuracy	0.99	***0.98***	0.30	0.05
	Bit rate	32	30	26	11
	Energy	2.3333	2.3329	2.3328	2.3325
Video encoding features matrix row 2	Accuracy	0.95	0.97	0.28	0.04
	Bit rate	30	29	24	10
	Energy	2.3329	2.3328	2.3321	2.3318
Multi-objective matrix row 1		0.68	0.96	0.27	0.04

Table 3 Another example illustrating the multi-objective function, considering only the accuracy [Weights: m = 1, n = 0.1, p = 0, q = 0, u = 0, and v = 0]

Row number	Specific video encoding features dimension	Entry 1	Entry 2	Entry 3	Entry 4
Video encoding features matrix row 1	Accuracy	0.99	***0.98***	0.30	0.05
Video encoding features matrix row 2	Accuracy	0.95	0.97	0.28	0.04
Multi-objective matrix row 1		0.68	0.96	0.27	0.04

$$g_H(A,R) = H(A+1,R+1) - H(A,R). \quad (2)$$

The multi-objective function can be applied to any specific case. For example, Table 3 illustrates the case when the bandwidth and energy are not of concern and only the accuracy is to be considered.

We then discuss how weights can be chosen and then introduce a method for finding the best video encoding features and analyze their time complexity.

3.3 Considerations in the multi-objective function weight settings

The variables (weights) m, p, u, n, q, and v can be set by system administrator or set as a function of system states. The weight 0 is the minimum preference and 1 is the highest preference (i.e. $\{weight \mid 0 \leq weight \leq 1\}$).

We suggest the following simple method to efficiently adjust the weights. The weights for the function accuracy can be set to the highest (1) or as a function of the recognized objects. On the other hand, video stream bit rate, and encoding power consumption should be set as a function of the available bandwidth and the remaining charge in the energy source, respectively. The bandwidth utilization metric can be *smoothed channel busy ratio*, which is indicator of how much is the channel busy during a given period of time, as specified in standards IEEE 802.3311p and SAE J2945.1 [31]. The derivative of the computer vision function accuracy, power consumption, and transmitted video bit rate are much less importance than their values. Thus their weights can be set zero or close to zero.

3.4 Proposed method for determining the optimal video encoding features for adapting video capturing and transmission

We can summarize the proposed method for finding the values of the controlling parameters that optimize the performance of the video communication in AI-based computer vision system as follows. During the deployment of the system, the administrator should capture a short video segment of the same conditions and environments. The system will encode this video segment utilizing the several control variables in for various pairing of frame size and quantization factor/scaling factor parameter. Next, it builds the video encoding features matrix by finding the CV function accuracy, consumed energy for video compression, and transmitted video bit rate for each system feature. Computer Vision function accuracy can be specified comparative to the original recording. The energy consumed can be calculated using an analytical model. Study [32] developed a model of the total consumed energy by a camera in the capture, encoding, and video transmission stages, based on frame size and quantization factor/scaling factor. After generating the video encoding features matrix, the receiving station applies the multi-objective function to produce the multi-objective matrix. At last, it looks for the highest values in the multi-objective matrix.

3.5 Time complexity of the proposed algorithm

Let us now analyze the time complexity incurred by the camera (sending station) for running the proposed method. For each mod (encoding controlling parameters pairing), we need to (1) encode the short video and then (2) decode it, (3) find various metrics (4) including running the CV algorithm for testing the accuracy, (5) approximating the energy from [33] (6) execute the multi-objective function, and then (7) find the maximum values of ABE_{MOF} output, which is the maximum of multi-objective matrix. This complexity is dominated by the encoding time.

The high time complexity of video encoding is basically a result of Motion Estimation (*ME*) and Rate Distortion Optimization (*RDO*) [32].

For each MacroBlock (*MB*), RDO performs a brute force search for the optimal block sizes and prediction modes, ME vectors, and other parameters that minimize the RDO cost J [32]. For a specified *MB*, the process proceeds as follows for the Macro-Block [32]: (i) Prediction calculation, (ii) Residual calculation, (iii) Residual encoding, (iv) Residual decoding, (v) Reconstructing, (vi) distortion computation, and (vii) cost computation J. This process is done iteratively for each pairing. After that, the setting with

the least cost will be selected for the Macro-Block. The complete process is repeated for each Macro-Block in the video frame. The encoding time complexity of one frame is basically the sum of the complexities of (i) intra- and inter-prediction, (ii) transformation, quantization, transformation inverse, quantization inverse, reconstruction, distortion and cost computations, and (iii) entropy encoding. Therefore, the encoding computation complexity X_e for F frames (averaging several frames in a second) is given by

$$X_e = X_{inter} + X_{intra} + X_{transquant} + X_{entropy}, \tag{3}$$

where X_{inter} is the the time complexity of inter-prediction multiplied by the ratio of inter-frames to all frames in F frames, X_{intra} is the time complexity of intra-prediction multiplied by the ratio of intra-frames to all frames in F frames, $X_{transquant}$ is the transformation, quantization, transformation inverse, quantization inverse for F frames, and $X_{entropy}$ is the entropy encoding time complexity for F frames.

We can express X_{inter} as the sum of integer ME complexity ($X_{integer}$) and fractional ME complexity ($X_{fractional}$):

$$X_{inter} = X_{integer} + X_{fractional}. \tag{4}$$

Based on the analysis in [32], the total of integer and fractional complexity (the whole inter-prediction) for a full search X_{inter} can be expressed by

$$X_{inter} \approx c_{integer} \times L \times R \times S^2 + c_{fractional} \times L, \tag{5}$$

and the intra complexity can be exxpressed by

$$X_{intra} = c_{intra} \times L, \tag{6}$$

where $c_{integer}$ and $c_{fractional}$ are constants, L is the pixels per second (pixel rate), R is the number of reference frames, and S is the search range. The bit rate r can be given as

$$r = c_{rate} \times \frac{L}{q^c}, \tag{7}$$

where q is the quantization parameter and c_{rate} is constant.

The transformation, quantization, transformation inverse, and quantization inverse for F frames can be expressed by

$$\begin{aligned} X_{traquant} &= c_{traquant} \times r = c_{traquant} \times c_{rate} \times L/q^c \\ &= c_{qnt} \times L/q^c, \end{aligned} \tag{8}$$

where $c_{traquant}$ and c_{rate} are constants, r is the bit rate.

The entropy complexity $X_{entropy}$ of a frame is linearly proportional to bit rate [34]. Based on Eq. (7), we can express it as

$$X_{entropy} = c_{entropy} \times r = c_{entropy} \times c_{rate} \times L/q^c = c_{bit} \times L/q^c. \tag{9}$$

Based on Eqs. (3), (5), (6), (7), (8), and (9) the complexity X_e of encoding F frames can be expressed as

$$\begin{aligned} X_e &= c_{integer} \times L \times R \times S^2 + (c_{fractional} + c_{intra}) \times L \\ &\quad + (c_{qnt} + c_{bit}) \times L/q^c \\ &= c_{integer} \times L \times R \times S^2 + c_L \times L + c_{Lq} \times L/q^c, \end{aligned} \tag{10}$$

$$= c_{all} \times L, \tag{11}$$

where $c_{integer}$, $c_{fractional}$, c_{intra}, c_{qnt}, c_{bit}, c_L, and c_{Lq}, c_{all} are constants, L is the pixels per second (pixel rate), R is the number of reference frames, S is the search range, and r is the bit rate. More detailed descriptions of these modes can be found in [29, 32].

Since the computational complexity of encoding a video is linear with the pixel rate (Eq. 11), We can formulate the total computational complexity (T_c) of the algorithm as:

$$T_c = O(N_{adaptations} \times N_{frames} \times R_x \times R_y), \tag{12}$$

where $N_{adaptations}$ is the video encoding features number, N_{frames} is the video frames number, and R_x and R_y are number of pixels in horizontal and vertical dimensions, respectively. The number of Several video encoding features sets is practically limited as a result of the limited practical quantization parameters and resolutions. Note that the overhead of searching for the optimal video encoding features is negligible compared to encoding times. based on this fact and the tight searching range, examination of the maximum value in the multi-objective array can purely be done by employing the brute-force algorithm. Furthermore, the video encoding features matrix can be refined based on learning from history.

4 Methodology of performance evaluation

4.1 Video data-sets used

Our dataset includes 300 real youtube collected videos and standard video segments of several qualities. Table 4

Table 4 Standard video segments properties

Sequence name	Duration (s)	Frame size	# Frames
Silent	10	CIF	300
Akiyo	10	CIF	300
Deadline	45.8	CIF	1374
SignIrene	18	CIF	540
vtc1nw	12	4SIF (VGA)	360

Table 5 Collected video dataset properties

Description	# Videos	Duration (s)	Frame Size	# Frames
Security	100	2857	QVGA	85,710
News	200	66096	QVGA	1,982,880
Total	300	68953	QVGA	2,068,590

summarizes the characteristics of the used video segments . We do not experiments with sequences of higher definition resolutions because Computer vision algorithms work well on low resolution and low video quality.

Table 5 contains a summary of the characteristics of the YouTube collected videos. The number of frames in all the YouTube videos are about two millions with several qualities and capturing environments from several Several cameras. For energy experiments, we use a twenty two minute video called "Baby Animal Songs by Kidsongs", which can be found on YouTube.

4.2 Creating simple videos with multiple qualities

We first create several video qualities by applying several adaptation (video encoding features) techniques individually. FFmpeg is used to convert videos to several frame sizes, frame rates, and video qualities. We encode each video using 10 or more Several lower frame size, at least 10 Several frame rates, and 10 or more Several target bit rates.

4.3 Combining adapting techniques

To Study pairing frame size and Quantization Factor (QF) video encoding features, we encode all the video segments in all the pairs of ten frame sizes and thirty one different *QF* settings. The first impact the resolution and the latter impact the *SNR*. We consider the average of accuracy, energy consumption, and bit rate of four Common Intermediate Format(*CIF*) sequences: Mother-daughter, News, Silent, and Foreman. We experiment with both the popular encoding standards H.264 and MPEG-4

4.4 Experimental setups

We experiment using three setups on two different platforms: a general purpose computer with FFmpeg (software encoder) and a security external camera (hardware encoder). Table 6 illustrates the three experimental setups. We follow the steps in [18] to measure and separate the consumed energy for each video communication stage (capturing, encoding, transmission).

In Experimental Setup III, we experiment with a CMOS networked surveillance camera (Vivotek IP713) with built-in WLAN (802.3311g), similar to Setup II.

We set the weights of ABE_{MOF} model for the illustrated results as follows: $m = 1.00$, $n = 0.10$, $p = 0.20$, $q = 0.20$, $u = 0.20$, and $v = 0.20$ for MPEG-4 in the two evaluated systems and $m = 1.00$, $n = 0.10$, $p = 0.10$, $q = 0.10$, $u = 0.10$, and $v = 0.10$ for H.264.

5 Results

We study the effectiveness of Several techniques for adapting the video communication: spatial (frame size), spatial with upscaling at the destination, temporal (frame rate), and signal-to-noise ratio (SNR). The latest is controlled by quantization factor (H.264) or scaling factor (MPEG-4). For experiments with individual video encoding features (techniques for adapting video transmission), in all the figures shown, we set the the parameter that is not under study to its maximum value. Unless otherwise noted, the mean value of the outcomes for video segments in Table 4 are illustrated.

5.1 The efficiency of upscaling frame sizes

To compare the efficiency of the super-resolution algorithms, Table 7 compares five upscaling (super-resolution) algorithms according to the face detection accuracy. The video segments in Table 4 are utilized, and the mean value of the detection accuracy is disclosed. Upscaling algorithms can improve the Computer Vision functions accuracy by a factor of twelve on average. Bicubic generally has the highest performance based on both detection

Table 6 Description of the experimental setups

Experimental setup	Platform	Input videos	Power consumption
I	Computer	Several datasets	Only encoding with actual measurements
II	Computer	Webcam	Aggregate with actual measurements
III	Surveillance camera	Surveillance camera	Aggregate with the power consumption model

Table 7 Upscaling algorithms comparison in detection accuracy percentage [H.264, experimental setup I]

Bit rate (Kbps)	None	Neighbor	Bilinear	Bicubic	Spline	Lanczos
70	0	37	40	40	41	41
220	0	86	85	88	88	88
260	18	91	90	90	90	90
Average	6	71	71	72	73	73

accuracy and computational complexity combined. Therefore, it is utilized in upscaling in all the paper experiments from this point on.

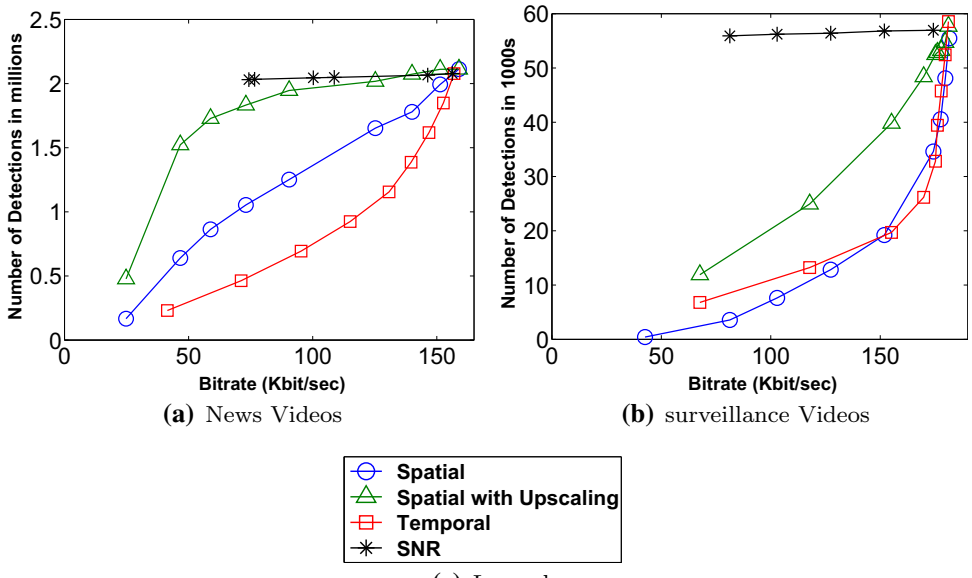
5.1.1 Video compression characterization for several codecs

Figure 3 characterises of various video stream video encoding features for standard video segments, and 300 YouTube collected videos. The detection accuracy is the the mean value of each category. These outcomes illustrate that the SNR and the spatial with upscaling framework display the maximum accuracy and encoding efficiency in the videos adaptions. Hence, changing the bit rate by varying the quantization parameter or the frame size has the slightest negative affect on the accuracy metric.

Figure 4 False Positive Index Analysis of Several techniques. Spatial with upscaling at the destination accomplishes mostly the maximum according to this metric.

5.1.2 Comparing video streams video encoding features in power consumption

So far, we analyzed various rate system feature in object/event detection accuracy. Let us now analyze them according to power consumption. The consumed energy in the capturing stage is linearly proportional to the total number of pixels in the video [29]. In this way, spatial and temporal video encoding features are anticipated to need less consumed energy for capturing than for SNR. In other hand, the consumed energy in the sending stage relies on the bitrate of the encoded video. The consumed energy in the encoding stage is the foremost critical among the three stages. Through actual experiments, we compared the broadcast of the three videos features according to total consumed energy in the capture, encoder, and transmission stages. Figure 5 demonstrate that spatial and temporal video encoding features lead to lower general consumed energy as they reduce the consumed energy in the encoding and capturing stage. consumed energy outcomes vary with implementation, But the common behavior will not alter as long as the gadgets utilizes dynamic voltage scaling.



(a) News Videos **(b)** surveillance Videos **(c)** Legend

Fig. 3 Characterization based on the chosen dataset [H.264, Experimental Setup I, Bicubic Interpolation Upscaling, Spatial where Spatial with upscaling and SNR performs the highest parameters for adapting video transmission]

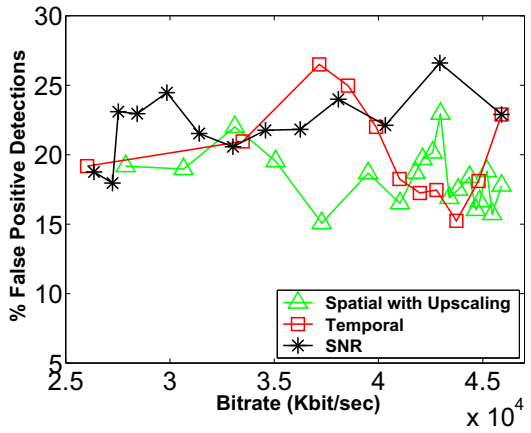


Fig. 4 False-Positive Index (*FPI*) [Experimental Setup I, Average Results for both MPEG-4 and H.264, Bicubic Interpolation Upscaling, the *FPI* is around 20%]

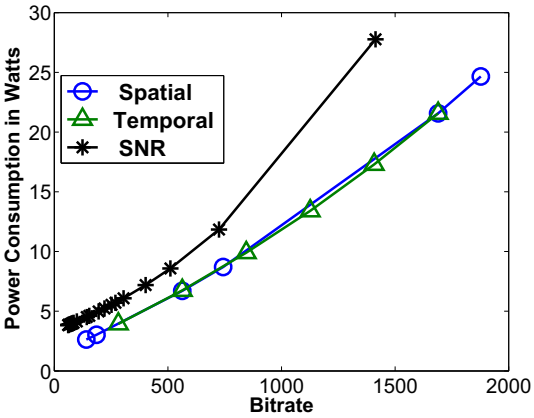


Fig. 5 Comparing stream video encoding features in aggregate power consumption of video capturing, encoding, and transmission [H.264, Experimental Setup II, Default Parameters: 768 × 480, 30 fps, and 22 QP, SNR consumes the most power than spatial or temporal adaptation]

5.2 SNR and spatial feature pairing outcomes

We refer to spatial system feature with upscaling simply as a spatial adaptation or spatial feature in this and the next subsection. Since spatial upscaling and SNR video encoding features performs the best, we examine how they are paired. Starting from this subsection, we represent *normalized* in the figures by shortly *N*. as in *N. Accuracy or N. Consumed Energy*. In addition, we represent normalized by shortly *M.N*. for the bit rate, for *manipulated normalized bit rate*, which is equal to $\frac{\sqrt{r}}{\sqrt{r_{max}}}$, where r and r_{max} represent the bit rate and maximum bit rate, respectively. This manipulation suits all bit rates of exponentially changing in a smaller range for clear representations in figures.

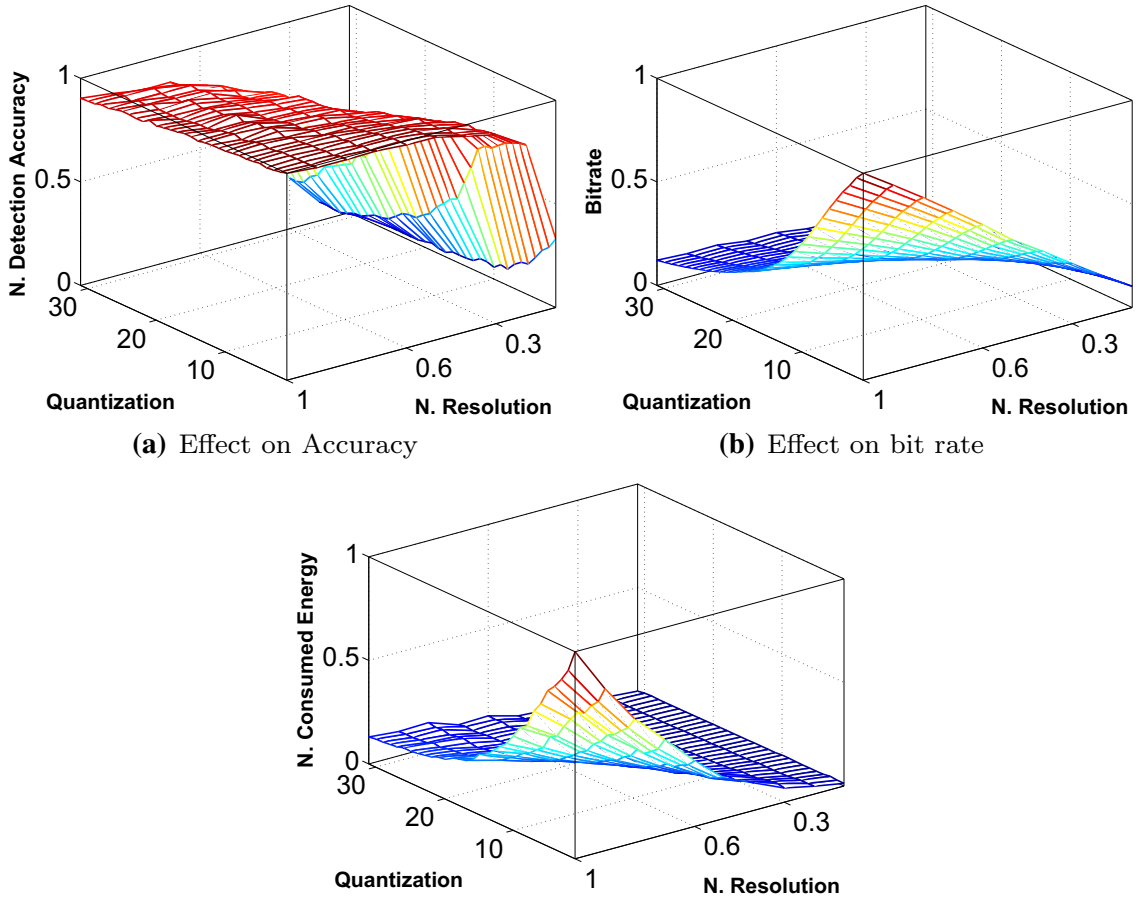
Figure 6 illustrates the accuracy metric, bit rate metric, and consumed energy as a function of both the quantization parameter and the resolution (frame size). Figure 6a show that the accuracy metric are not altered significantly by varying the *QP*, but it changes very fast by varying the frame size below half the original image. However, in Fig. 6b, the bit slope are very steep and the change is very fast for low *QP*. In Fig. 6c, the consumed energy is reduced when lowering the *QP* and/or the frame size by up to half. Considering these outcomes, we must prevent low *QP* setting and frame sizes less than half that of the original sizes because of the high drop in bit rate and accuracy metric for low *QPs* and frame sizes, respectively.

5.3 Analysis of the outcomes of the proposed optimization function

Now let's explain the proposed ABE_{MOF} which is developed to find the optimal encoding process in terms of accuracy, bandwidth requirements, and power consumption. The process involves combining multiple features of the video encoders. Figure 1 shows the ABE multi-objective matrices which is the outcome of ABE_{MOF} model for MPEG-4 and H.264 when varying both the frame size and Signal-to-Noise-Ratio (SNR) (i.e., *QP* for H.264 or *SF* for MPEG-4).

We study the impact of applying ABE_{MOF} for SNR and the resolution video encoding features, and their pairing. Figures 8 and 9 illustrate the impact of three SNR choice tactics at Several resolutions for MPEG-4 and H.264, respectively. The illustrated methodologies are the *Maximum SNR*, the *Minimum SNR*, and the *Model SNR*. The first two strategies refer to choosing the maximum and minimum SNR settings, respectively. The maximum settings related to the lower values of the *QP* or *SF*, which produces the maximum quality and accuracy metrics. Therefore, the first methodology outcome is the maximum accuracy metric, while the second methodology outcomes is the minimum bit rate and consumed energy metrics. However, Model *SNR* sets the *QP* or *SF* according to ABE_{MOF} for the given Frame size.

Figures 10 and 11 analyse the impact of selecting three different frame sizes tactics: *Minimum Resolution*, *Maximum Resolution*, and *Model Resolution*. The tactics are compared at various SNR settings. The first two refer to choosing the minimum and maximum frame sizes in the studied epoch, respectively. Consequently, the first strategy produces the lowest bit rate and energy consumption, while the second produces the maximum accuracy. Model Resolution sets the frame size, however, based on ABE_{MOF} model for the given SNR settings. Figure 10 shows the impact on accuracy metric and consumed energy separately using H.264, where Fig. 11 shows the total energy-accuracy characteristics for both MPEG-4 and H.264. These outcomes illustrate that by using ABE_{MOF}, we can

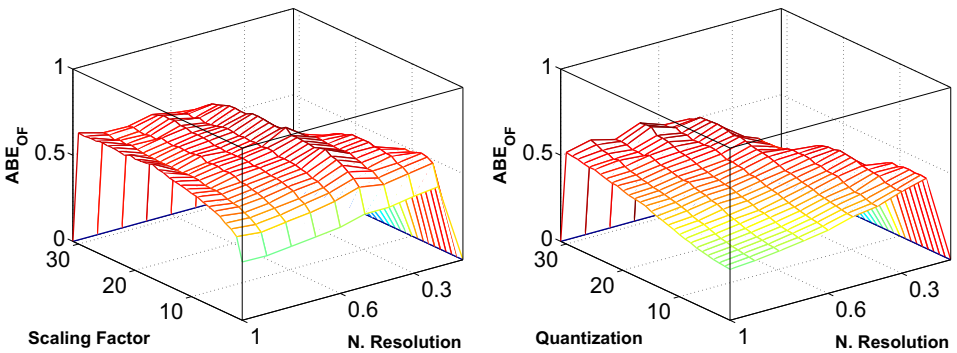


(a) Effect on Accuracy **(b)** Effect on bit rate

(c) Effect on Consumed Energy

Fig. 6 Comparison of sets of several resolutions and SNR [H.264, experimental setup I, QP of 20 and 50% of the resolutions keeps the accuracy unaffected but reduce the energy and bit rate to less that 10% of the original]

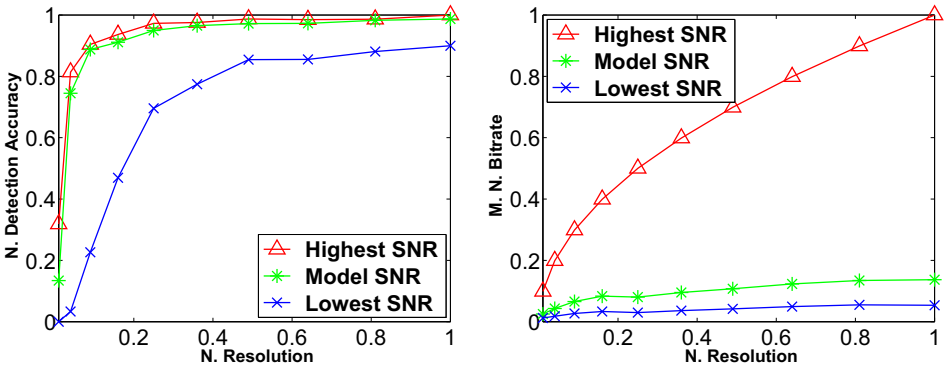
Fig. 7 Accuracy-bitrate-energy tradeoff: comparison of sets of several Resolutions and SNR [Experimental Setup I, This Figure of the Objective Function for selected parameters, support Fig. 7; QP of 20 and 50% of the resolutions shows the Objective Function (ABE_{MOF} or ABE_{OF}), has maximum values which optimize the bit rate and energy consumption for the highest accuracy metric]



(a) MPEG-4 **(b)** H.264

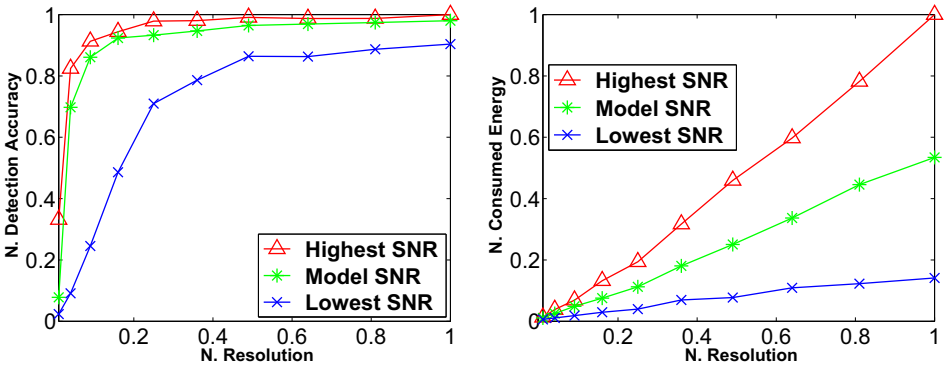
accomplish an accuracy close to the maximum, at the same time we extremely reduce the bit rate and consumed energy.

Now we will analyze the effectiveness of applying ABE_{MOF} on combining SNR with spatial video encoding features (including upscaling). Figures 12 and 13 compare the impact of three pairs of both resolution and SNR for MPEG-4 and H.264, respectively. The analyzed tactics are *Minimum Pairing*, *Maximum Pairing*, and *Model Pairing*. Setting one and two are to generate the minimum and maximum videos qualities, respectively. The parameters pairing settings is based on ABE_{MOF} model. The figures show both the bit rate-accuracy and energy-accuracy characteristics. The results identify three separate classes:



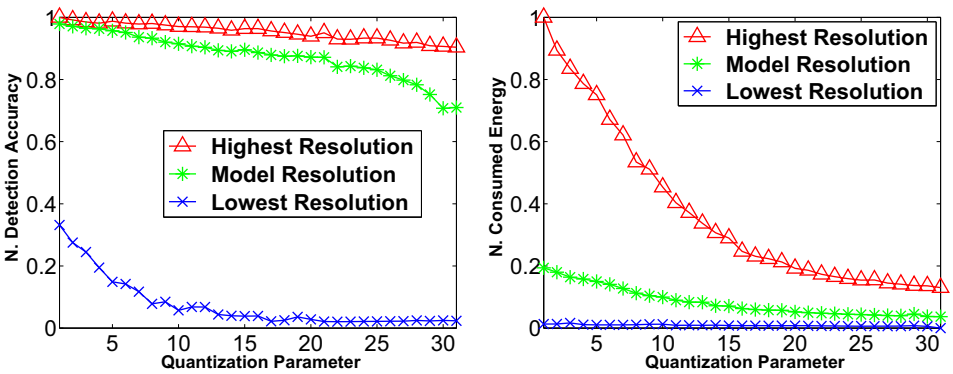
(a) Effect on Accuracy **(b)** Effect on Bitrate

Fig. 8 Analyzing SNR epoch at several resolutions [MPEG-4, experimental setup I, the detection accuracy is almost identical to the highest (smallest values of the scaling factor), The bit rate is slightly higher than the minimum]



(a) Impact on Accuracy Metric **(b)** Impact on Consumed Energy metric

Fig. 9 Analyzing SNR epoch at several resolutions [H.264, experimental setup I, The *DetectionAccuracy* is slightly less than the highest (smallest values of the *QP*), The *ConsumedEnergy* is in the middle between the maximum and the minimum]



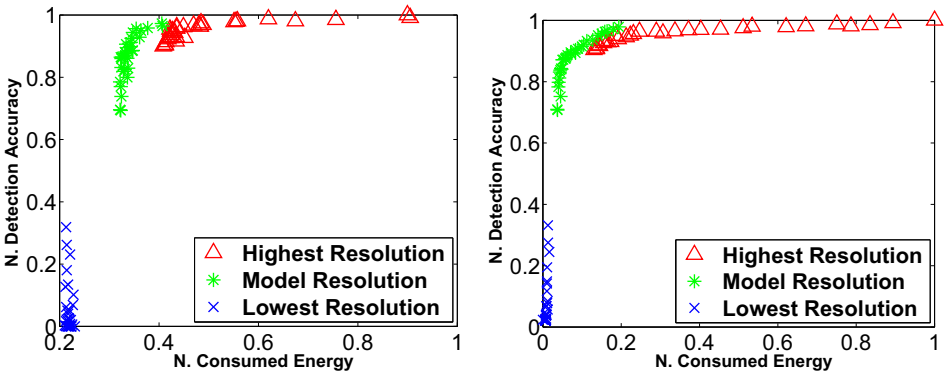
(a) Effect on Accuracy **(b)** Effect on Consumed Energy

Fig. 10 Analyzing resolution range at several SNR [H.264, setup I, the accuracy metric is slightly below the maximum (smallest values of the resolutions parameter), The energy consumption is slightly higher than the minimum]

the first has maximum accuracy metrics, the second has minimum bit rates and consumed energy, and the optimal is the result of ABE_{MOF}. By pairing SNR and frame size video encoding features in H.264, we able to save 80% (60% for MPEG-4) of the energy requirements and 98% (99% for MPEG-4) of the bit rate, while keeping the accuracy metric without reduction more than 10% (5% for MPEG-4) of the maximum.

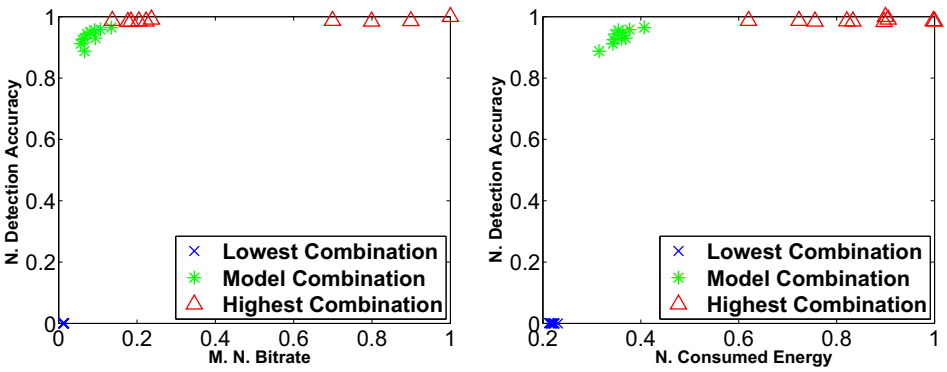
Let us now compare the effectiveness of applying ABE_{MOF} for SNR system feature, frame size feature, and their pairing. Figure 14 shows the impact of Model SNR, Model Resolution (frame size), and Model pairing in bit rate-accuracy and bit rate-energy characteristics. These results illustrate the advantage of pairing video encoding features in an optimal way.

At last, we explain the outcome under Experimental Setup III, which utilizes the security camera. Figure 15a



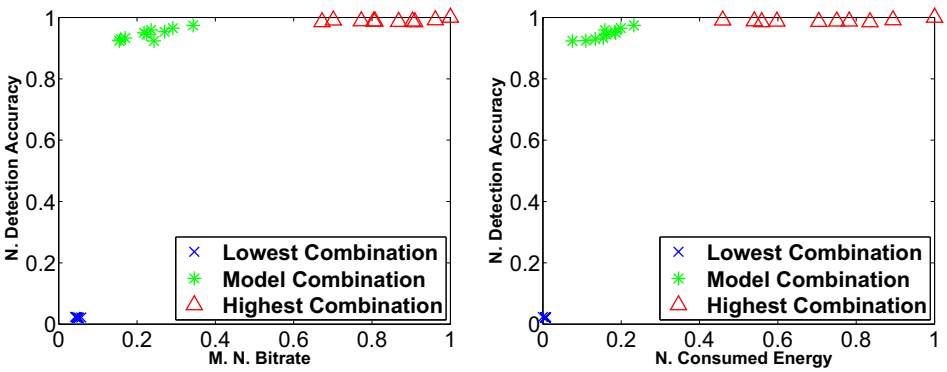
(a) MPEG-4 **(b)** H.264

Fig. 11 Energy-accuracy tradeoff: comparing resolution selection at several SNR [experimental setup I, these outcomes illustrate that by using ABE_{MOF}, we can accomplish an accuracy close to the maximum, at the same time we extremely reduce the bit rate and consumed energy.]



(a) Bitrate-Accuracy Tradeoff **(b)** Energy-Accuracy Tradeoff

Fig. 12 Comparison of sets of several resolutions and SNR [MPEG-4, experimental setup I, by pairing SNR and frame size video encoding features in MPEG-4, we able to save 60% of the energy requirements and 99% of bit rate, while keeping the accuracy metric without reduction more than 5% of the maximum.]



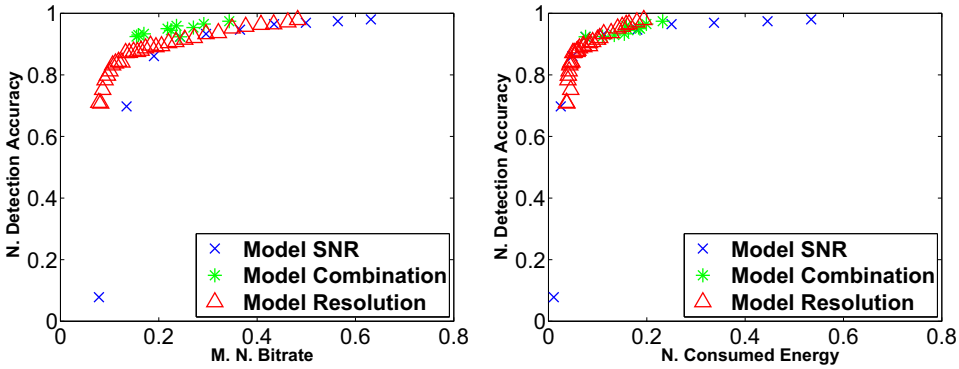
(a) Bitrate-Accuracy Tradeoff **(b)** Energy-Accuracy Tradeoff

Fig. 13 Comparison of sets of several resolutions and SNR [H.264, experimental setup I, By pairing SNR and frame size video encoding features in H.264, we able to save 80% of the energy requirements and 98% of the bit rate, while keeping the accuracy metric without reduction more than 10% of the maximum.]

analyzes the impact of Minimum Resolution, Model Resolution, and Maximum Resolution at several SNR settings, while Fig. 15b analyzes the impact of Minimum Pairing, Model Pairing, and Maximum Pairing. These results demonstrate similar behavior as those under Experimental Setup I, with the most distinction is in consumed energy.

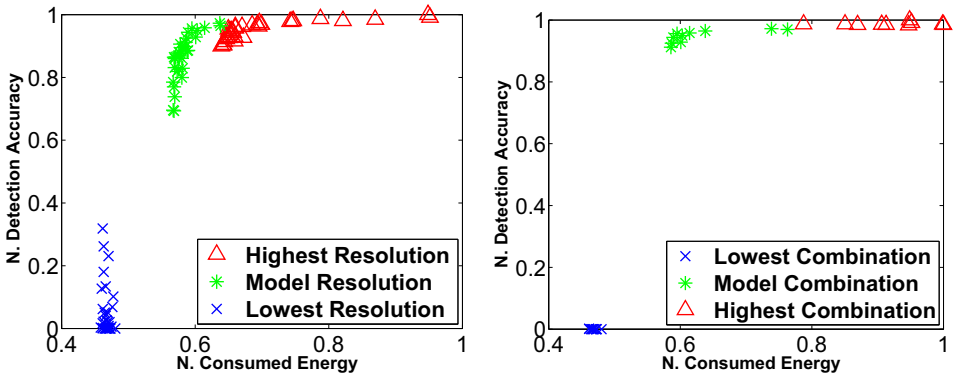
6 Conclusions

We analyzed the bitrate-accuracy and bitrate-power consumption characteristics of four video encoding features for adapting video capturing and transmission [spatial (frame size), spatial with upscaling (downscale and upscale of frame size), temporal (frame rate), and SNR (quality)] by performing real-world experiments using H.264 and MPEG-4 codecs, taking into account special research video



(a) Bitrate-Accuracy Tradeoff **(b)** Energy-Accuracy Tradeoff

Fig. 14 Analyzing model (ABE_{MOF}) SNR, resolution, and pairings (pairs) [H.264, experimental setup I, this figure analyzes the impact of applying ABE_{MOF} for SNR system feature, frame size feature, and their pairing]



(a) Resolution **(b)** pairing

Fig. 15 Comparing several video encoding features [MPEG-4, setup III, Fig. 15a compares the effectiveness of minimum resolution, model resolution, and maximum resolution at several SNR settings, whereas Fig. 15b compares the effectiveness of minimum pairing, model pairing, and maximum pairing]

segments and a collected dataset from YouTube of three hundred videos with over one million frames.

The analysis of the outcome of the experiments shows that the SNR feature is the best overall feature, followed by spatial and upscaling the frame sizes at the destination.

Furthermore, we introduced a multi-objective function and optimization tactic to determine the adaptive or adaptive mix that provides optimal performance according to accuracy of function detection, encoding energy consumption, and bandwidth requirements.

The Multi-objective function supports higher computer vision function accuracy, lower bit rate, and lower power consumption. Furthermore, the multi-objective function supports the setup with the next decrease in CV function accuracy, a smaller next decrease in bit rate, and the next decrease in consumed energy.

The analysis of the experiment's outcome suggests that we must stay a way from high numbers of *QP* and less than half he primary resolution, because of the high reduction in performance.

The multi-objective function can be used if there is one system feature or numerous video encoding features. For a single system feature, the analysis of the experiment outcomes shows that the multi-objective function reaches optimal accuracy, bitrate, consumed energy. For several video encoding features, the analysis of the outcome of the experiments show the advantage of combining video encoding features in pairs. For example, by pairing signal-to-noise ratio (SNR) and spatial feature in H.264, the consumed energy can be reduced to less than 20% of the maximum, and the bit rate can be reduced to less than 2% while maintaining the CV function accuracy metric within 10% of the maximum.

Toward optimal performance in the future work, we will utilize deep reinforcement learning to set the weights dynamically for the computer vision function accuracy, bandwidth requirements, power consumption, and their rates of change, respectively. In this paper, we set these weights manually based on our knowledge of the system. Again, setting the weights should be done offline during the calibration process. In addition, we will include more encoders for testing the system performance based on the proposed optimization strategy and multi-objective function.

References

1. Nguyen, M.T., Truong, L.H., Tran, T.T., Chien, C.-F.: Artificial intelligence based data processing algorithm for video surveillance to empower industry 3.5. Comput. Ind. Eng. **148**, 106671 (2020)
2. Tiefenau, C., Häring, M., Krombholz, K., von Zezschwitz, E.: Security, availability, and multiple information sources: Exploring update behavior of system administrators. In *Sixteenth Symposium on Usable Privacy and Security ({SOUPS }2020)*, pp. 239–258 (2020)
3. Alsmirat, M., Sarhan, N.J.: Intelligent optimization for automated video surveillance at the edge: A cross-layer approach. Simul. Model. Pract. Theory **105**, 102171 (2020)
4. Mama, C., Noureddine, B., Benaissa, B.: Control of variable reluctance machine (8/6) by artificial intelligence techniques. Int. J. Electr. Comput. Eng. (2088–8708) **10**, 2 (2020)
5. Korshunov, P., Ooi, W.T.: Video quality for face detection, recognition, and tracking. ACM Trans. Multimedia Comput. Commun. Appl. **7**(3), 1–21 (2011)
6. Alsmirat, M., Sarhan, N. J.: Cross-layer optimization for automated video surveillance. In *IEEE International Symposium on Multimedia (ISM)*, pp. 243–246, December (2016)
7. Wang, Z., Chen, J., Hoi, S. C. H.: Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, arXiv:1902.06068 (2020)
8. Son, S., Lee, J., Nah, S., Timofte, R., Lee, K. M., Liu, Y., Xie, L., Siyao, L., Sun, W., Qiao, Y., et al.: Aim 2020 challenge on video temporal super-resolution. In *European Conference on Computer Vision*, pp. 23–40. Springer (2020)
9. Sharrab, Y. O., Sarhan, N. J.: Accuracy and power consumption tradeoffs in video rate adaptation for computer vision applications. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 410–415 (2012)
10. Javed, O., Shah, M.: Tracking and object classification for automated surveillance. In *Proceedings of the European Conference on Computer Vision-Part IV*, pp. 343–357, (2002)
11. Yuan, X., Sun, Z., Varol, Y., Bebis, G.: A distributed visual surveillance system. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, p. 199 (2003)
12. Niu, W., Long, J., Han, D., Wang, Y.-F.: Human activity detection and recognition for video surveillance. In *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, pp. 719–722 (2004)
13. Kim, J., Wang, Y., Chang, S.: Content-adaptive utility-based video adaptation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 281–284 (2003)
14. Hamandi, H. R., Sarhan, N. J.: Novel analytical models of face recognition accuracy in terms of video capturing and encoding parameters. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE (2020)
15. Barzigar, N., Roozgard, A., Verma, P., Cheng, S.: A video super-resolution framework using SCoBeP. IEEE Trans. Circuits Syst. Video Technol. **26**(2), 264–277 (2016)
16. Georgis, G., Lentaris, G., Reisis, D.: Reduced complexity superresolution for low-bitrate video compression. IEEE Trans. Circ. Syst. Video Technol. **26**(2), 332–345 (2016)
17. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. Cengage Learning, Boston (2014)
18. Sharrab, Y.: Video stream adaptation in computer vision systems. digitalcommons.wayne.ed (2017)
19. Sarif, B.A.B., Pourazad, M. T., Nasiopoulos, P., Leung, V.: Encoding and communication energy consumption trade-off in H.264/AVC based video sensor network. In *Proceedings of the IEEE International Symposium and Workshops on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–6, (2013)
20. Brown, T. X.: Low power wireless communication via reinforcement learning. In: Advances in Neural Information Processing Systems, pp. 893–899, Citeseer (2000)
21. Azar, A. T.: Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), vol. 1153. Springer Nature (2020)
22. Meske, C., Bunde, E.: Using explainable artificial intelligence to increase trust in computer vision. arXiv:2002.01543 (2020)
23. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, (2017)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, (2016)
25. Liu, W., Wen, Y., Zhiding, Y., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, p. 1 (2017)
26. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. (2018) CoRR (arXiv:abs/1801.04264)
27. Esterle, L., Lewis, P. R.: Online multi-object k-coverage with mobile smart cameras. In *Proceedings of the 11th International Conference on Distributed Smart Cameras (ICDSC)*, pp. 107–112. ACM, (2017)
28. Sharrab, Y. O., Sarhan, N. J.: Detailed comparative analysis of vp8 and h. 264. In *2012 IEEE International Symposium on Multimedia*, pp. 133–140. IEEE, (2012)
29. Sharrab, Y. O., Sarhan, N. J.: Aggregate power consumption modeling of live video streaming systems. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pp. 60–71 (2013)
30. Caramia, M., Dell'Olmo, P.: Multi-objective optimization. In: Multi-objective management in freight logistics. Springer, pp. 21–51 (2020)
31. SAE International: On-board system requirements for V2V safety communications. *Standard J2945/I_201603*, March (2016)
32. Sharrab, Y.O., Sarhan, N.J.: Modeling and analysis of power consumption in live video streaming systems. ACM Trans. Multimedia Comput. Commun. Appl. **13**(4), 1–25 (2017)
33. Sharrab, Y.O., Alsmirat, M., Hawashin, B., Sarhan, N.: Machine learning-based energy consumption modeling and comparing of H. 264 and google vp8 encoders. Int. J. Electr. Comput. Eng. **11**(2), 2088–8708 (2021)
34. He, Z., Liang, Y., Chen, L., Ahmad, I., Dapeng, W.: Power-rate-distortion analysis for wireless video communication under energy constraints. IEEE Trans. Circ. Syst. Video Technol. **15**(5), 645–658 (2005)

Yousef O. Sharrab is a professor and an expert in artificial intelligence, surveillance systems, and intelligent transportation systems. Dr. Sharrab has a Ph.D. in computer engineering and artificial intelligence from Wayne State University, Michigan, USA. His research experience and interest include Deep Learning and Computer Vision in the Domain of Automated Surveillance System, Network Security, Intelligent Transportation Systems, and Education. Dr. Sharrab worked for General Motors in the R&D Global Center for a USA national safety project and has over 20 years' experience in technology-based systems for governments and private sectors.

Izzat Alsmadi is an Associate Professor in the department of computing and cyber security at the Texas A&M, San Antonio. He has his master and PhD in Software Engineering from North Dakota State University in 2006 and 2008. He has more than 100 conference and journal publications. His research interests include: Cyber intelligence, Cyber security, Software security, software engineering, software testing, social networks and software defined networking. He is lead author, editor in several books including: Springer The NICE Cyber Security Framework Cyber Security Intelligence and Analytics, 2019, Practical Information Security: A Competency-Based Education Course, 2018, Information Fusion for Cyber-Security Analytics (Studies in Computational Intelligence), 2016. The author is also a member of The National Initiative for Cybersecurity Education (NICE) group, which meets frequently to discuss enhancements on cyber security education at the national level.

Nabil J. Sarhan is an Associate Professor of Electrical and Computer Engineering at Wayne State University and the Director of Wayne State Deep Learning Research Laboratory. He served as the Graduate Program Director of Electrical and Computer Engineering and the Chair of the College of Engineering Faculty Assembly.

Authors and Affiliations

Yousef O. Sharrab[1,2] · Izzat Alsmadi[3] · Nabil J. Sarhan[4]

✉ Izzat Alsmadi
ialsmadi@tamusa.edu

Yousef O. Sharrab
yousef.sharrab@wayne.edu

Nabil J. Sarhan
nabil.sarhan@wayne.edu

[1] Deep Learning Lab, Wayne State University, Detroit, USA

[2] Artificial Intelligence Department, Irbid National University & Hashemite University, Irbid & Zarqa, Jordan

[3] Department of Computing & Cyber Security, Texas A&M, San Antonio, USA

[4] Deep Learning Lab & Electrical & Computer Engineering Department, Wayne State University, Detroit, USA