The Impact of Video Transcoding Parameters on Event Detection for Surveillance Systems

Emmanouil Kafetzakis, Christos Xilouris
and Michail Alexandros Kourtis
Institute of Informatics & Telecommunications
National Center of Scientific Research "Demokritos"
P.O. Box 60228, GR-15310, Ag. Paraskevi, Greece
Email: mkafetz@iit.demokritos.gr
cxilouris@iit.demokritos.gr
akis.kourtis@iit.demokritos.gr

Marcos Nieto
Vicomtech-IK4
20009, San Sebastian, Spain
Email: mnieto@vicomtech.org

Iveel Jargalsaikhan and Suzanne Little
CLARITY Centre
for Sensor Web Technology
Dublin City University, Ireland
Email: iveel.jargalsaikhan@dcu.ie
suzanne.little@dcu.ie

Abstract—**The process of transcoding videos apart from being intensive, can also be a rather complex procedure. The complexity refers to the choice of appropriate parameters at the transcoding engine, towards decreasing video sizes, transcoding times and network bandwidth without degrading video quality beyond some threshold that event detectors lose their accuracy. The paper explains the need for transcoding, and then studies different video quality metrics. Commonly used algorithms for motion and person detection are briefly described, with emphasis in investigating the optimum transcoding configuration parameters. The analysis of the experimental results reveals that the existing video quality metrics are not suitable for automated systems, and that the detection of persons is affected by the reduction of bit rate (more blockiness effect) and resolution (less information), while motion detection is more sensitive to frame rate.**

I. Introduction

The existing Closed-Circuit Television (CCTV) infrastructures and surveillance video systems are not actually fully exploited. Scanning massive amounts of recorded video of different formats in order to locate a specific segment based on semantic descriptions remains a non-automated task, mainly performed by humans. For example, the Chicago's video surveillance camera system has more than ten thousand cameras [1] connected to a common storage system. The SAVASA project [2] aims to develop a standard-based video archive search platform that allows authorised users to query over various remote and non-interoperable video archives of CCTV footage from geographically diverse locations. At the core of the search interface is the application of algorithms for person/object detection and tracking, activity detection and scenario recognition.

In most platforms that aim to the decoupling of CCTV and Video Archive installations, video transcoding performs two fundamental operations: a) provide video format conversion to enable a unified data interface, and b) perform compression to facilitate the video annotation and storage. Video analysis and watermarking can be performed more easily when videos are compressed beforehand. In this paper, conversions between MPEG-2 coding standard [3] to H264/MPEG-4 Advanced Video Coding (AVC) [4] standard are performed.

The key disadvantage of transcoding is that more frequently it is a lossy process, introducing image artifacts (e.g., twisted or deformed images) and resulting in decreased video quality output. However, for large scale CCTV installations, the transcoding process is inevitable due to the diversity of CCTV cameras and their recording capabilities. In fact, since typical End-Users do not constantly observe all video streams but only rare suspicious events [5], they do not have high-quality video requirements. In future surveillance systems, the videos will be mainly transmitted for the automated video analysis algorithms, with the minimum acceptable quality for increasing the scalability of the CCTV systems.

Nevertheless, the video quality should not be degraded beyond some threshold so that event detectors do not lose their accuracy. In this direction, we measure the video quality deterioration in terms of Peak Signal to Noise Ratio (PSNR) [6] and Frame Rate Structural SIMularity (SSIM) [7] full reference metrics. Although these metrics have been widely used as video quality indicators, this paper brings out that they are not suitable to demonstrate the degree that event detectors are affected by the compression. In some cases it is observed that the apparently reduced video quality gives better results for the cases of motion and person detection. Even though this is initially somewhat strange, it can be explained by the fact that the existing video quality metrics simulate human perception that may be different from computer vision.

The main contribution of this paper is the study of the accuracy of common event detectors in relation to the input video quality. The adopted motion detection algorithm is based on descriptors for motion trajectories, which are calculated using salience points identified by Harris Corner detectors [8] and tracked using the Kanade-Lucas-Tomasi (KLT) algorithm [9], [10]. Trajectories are described using four descriptors, and then they are classified via a trained Support Vector Machine (SVM). Persons are detected using Histogram of Oriented Gradients (HOG) descriptors [11] and tracked via Rao-Blackwellized Data Association Particle Filter [12]. Other event detection algorithms, used in SAVASA project, can be found in [13].

The lowest video quality allowing humans to perform recognition of natural image contents is studied in [14]. From computer vision perspective, the most relevant work to ours is [15] which demonstrates also that the face detection algorithms show almost no decrease in accuracy until the input video is reduced to a certain critical quality. Our work investigates the critical quality for full-body person detection and pointing detection using an open data set.

The rest of the paper is organized as follows: Section II presents the transcoding parameters under investigation, while Section III briefly reviews common video quality metrics. The adopted event detection methodology is described in Section IV. The input video quality measurements and the evaluation of detectors in relation to the transcoding parameters are included in Section V. Section VI concludes the paper.

II. Transcoding Parameters

The following transcoding parameters affect input video quality and they need to be considered for automated event detection.

A. Bit Rate

In computer vision area, bit rate refers to the amount of detail that is processed in a predefined time duration. Bit rates can be classified into two main categories: Variable Bit Rate (VBR) and Constant Bit Rate (CBR) encodings. VBR permits a higher bit rate to be allocated to the more high motion scenes and to the complex segments of videos, and a less rate to be allocated to less complex segments. More specifically, when there is little or no motion on the scene, the encoder decreases the bit rate to minimum bit rate, while when the motion is prevalent it is increased to the maximum allowed. This flexibility allows smaller overall file sizes without serious compromises in the quality of the video. Averaging the instant rates, the mean video bit rate value is calculated.

Resource allocation is easier with CBR, since bit rate is flat and thus predictable. This characteristic comes at the price of encoding efficiency; usually resulting in a larger file. CBR is suitable for streaming multimedia content on limited capacity networks, where multiplexing gain is limited. In contrast, CBR would not be the optimal choice for minimum storage space as it would not produce enough data for complex segments (leading to low quality), while sacrificing data on simple sections. In order to have a broad picture of bit rates in CCTV systems, note that one camera might produce between 100 kbps and 2 Mbps.

B. Video Resolution

Resolution is a measurement of the number of pixels in a frame. Each pixel is a piece of a puzzle which by itself it might not mean much, but when combined with other pixels it becomes a critical piece of information that helps to comprehend a larger visual story. As more pixels exist to the frame (thereby increasing its resolution), the image gets sharper and more detailed.

Typically, the resolution is expressed as frame length times height (both in pixels). Common resolutions for CCTV IP cameras are the CIF (640x480), the 4CIF (704x480) and the D1(720x480). The resolution 1280x720 is the minimum that is called High Definition (HD). There is also 1920x1080 resolution, which is sometimes referred to as full HD.

C. Frame Rate

This parameter specifies the number of frames that are generated/transmitted during a time unit – the higher frame rate, the smoother video is. Due to bandwidth and storage restrictions, CCTV systems use in practice frame rates between 5-15 frames per second (fps), which are sufficient in general. Lower frame rates are used in premises with little movement and in applications like crowd control, while higher rates (e.g., 25 fps) to monitor the behaviour of individuals in a realistic manner.

III. Video Quality Assessment

Simultaneously with the transcoding, a lossy video encoding technique can be applied to reduce the bandwidth needed to transmit or store video data, having as result the degradation of the quality. For this reason, it is crucial for an automated event detection surveillance system to be able to realize and quantify the video quality degradations, so that it can maintain and control the quality of the video data. Over the last years, emphasis has been put on developing various methods and techniques for evaluating the perceived quality of video content by human observers. These methods have not been designed for CCTV task-based applications, but mainly for entertainment. From the computer vision perspective, the fundamental measure of video quality is the success rate of recognition tasks. In this context, new initiatives are trying to address the lack of suitable metrics [16]. Since all these works are in a very early stage, we review here only well established video quality metrics, categorized into two broad classes: the subjective and the objective ones.

A. Subjective quality

The subjective test methods involve an audience of people, who watch a video sequence and score its quality as perceived by them, under specific and controlled watching conditions. The Mean Opinion Score (MOS) is regarded as the most reliable method of quality measurement and it has been applied on the most known subjective techniques: the Single Stimulus Continue Quality Evaluation (SSCQE) and the Double Stimulus Continue Quality Evaluation (DSCQE) [17]–[19]. However, the MOS method is usually inconvenient due to the fact that the preparation and execution of subjective tests is costly and time consuming.

Subjective test methods are also described in International Telecommunication Union-Radio (ITU-R) Rec. T.500-11 [20] and ITU-T Rec. P.910 [21], suggesting specific viewing conditions, criteria for observers and test material selection, assessment procedure description and statistical analysis methods. The ITU Rec. T.500-11 described subjective methods that are specialized for television applications, whereas ITU-T Rec. P.910 is intended for multimedia applications.

B. Objective quality

Since subjective tests are costly and time consuming, a lot of effort has been focused on developing cheaper, faster and easier applicable objective evaluation methods. These techniques successfully emulate the subjective quality assessment results, based on criteria and metrics that can be measured objectively. The objective methods are classified, according to

Quality Metric	Mathematical Complexity	Correlation with Subj. Methods
PSNR	Simple	Poor
SSIM	Complex	Fairly good

TABLE I. COMPARISON OF OBJECTIVE METRICS.

the availability of the original video signal, which is considered to be of high quality.

The majority of the proposed objective methods in the literature require the undistorted source video sequence as a reference entity in the quality evaluation process, and due to this are characterized as Full Reference Methods (see, e.g., [22], [23]). These methods are based on an Error Sensitivity framework with most widely used metrics the Peak Signal to Noise Ratio (PSNR) and the Mean Square Error (MSE).

Despite several objective video quality models have been developed in the past two decades, PSNR continues to be the most popular evaluation of the quality difference among videos, i.e.,

$$\text{PSNR} = 10 \log_{10} \frac{L^2}{\text{MSE}}, \qquad (1)$$

where L denotes the dynamic range of pixel values (equal to 255 for 8 bits/pixel monotonic signal). The MSE is defined by

$$\text{MSE} = \frac{\sum_{i=1}^{N} (x_i - y_i)^2}{N}, \qquad (2)$$

where N denotes the number of pixels, and x_i, y_i the i^{th} pixel in original, distorted frame, respectively.

The Frame Rate Structural SIMularity (SSIM) metric is an other objective metric which benchmarks the encoding efficiency of different block sizes in relevance to the spatio temporal activity level of the video content. SSIM is a Frame Rate metric for measuring the structural similarity between two image sequences, exploiting the general principle that the main function of the human visual system is the extraction of structural information from the viewing field and it is not specialized in extracting the errors. If x and y are two video frames,

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \qquad (3)$$

where μ_x, μ_y are the mean value of x and y, σ_x, σ_y, σ_{xy} are the variances of x, y and the covariance of x and y, respectively. The constants C_1 and C_2 are defined as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where L is the dynamic pixel range and $K_1 = 0.01$ and $K_2 = 0.03$, respectively [24]. The value of SSIM ranges between -1 and 1, and gets the best value of 1 if $x_i = y_i$ for all values of i.

A comparison of the objective metrics PSNR and SSIM is presented in Table I. It can be seen that the SSIM gives the most reliable result. However, the computational complexity of PSNR makes it ideal to apply in real-time applications. For these reasons, PSNR and SSIM are both used as baseline video quality metrics for the transcoding procedure.

IV. PERSON AND MOTION DETECTION

This section outlines the two classifiers used to identify video segments that show one of the two following events: a) Pointing and b) Person-Walks.

A. *Pointing Recognition Using Motion Trajectory*

To represent motion, we have used salience points for capturing the motion trajectory. This low-level feature is then described by four different descriptors. Firstly, in order to facilitate motion trajectory extraction, a background subtraction algorithm [25] to detect foreground regions has been applied. This stage reduces computational complexity and increases the accuracy of point tracking by reducing the searchable area. Salience points are located within the foreground regions by Harris Corner Detector [8] and are tracked over video sequences using Kanade-Lucas-Tomasi (KLT) algorithm [9], [10]. In the experiments, we have observed that longer salience points trajectories are likely to be erroneous. Therefore, we have empirically set the maximum trajectory length to be fifteen frames.

For the motion trajectory description, we adopted the approach in [26] to describe the trajectory features. For each trajectory, we calculated four descriptors to capture the different aspects of motion trajectory. Among the existing descriptors, HOG/HOF [27] has shown to give excellent results on a variety of datasets [28]. Therefore, HOG/HOF is computed along our trajectories. HOG (Histogram of Oriented Gradient) [11] captures the local appearance around the trajectories, whereas HOF (Histogram of Optical Flow) captures the local motion [26]. Additionally, MBH (Motion Boundary Histogram), proposed in [29], and TD (Trajectory Descriptor) [26] are computed in order to represent the relative motion and trajectory shape.

In order to represent the video scene, we have built a Bag-of-Features (BoF) model based on the four descriptors. This step requires the construction of a visual vocabulary. In this direction, we clustered a subset of 250,000 descriptors sampled from the training videos with the k-means algorithm applied for each descriptor. The number of clusters is set to k=4000, which has shown empirically to give good results in [27]. The BoF representation then assigns each descriptor to the closest vocabulary word in Euclidean distance and computes the co-occurrence histogram over the video sub-sequence.

Finally, we have used a non-linear Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel for the classification. Using the cross-validation technique, we have empirically found the parameters of cost (32) and gamma (10^{-5}) of the kernel. In order to represent the video frame, we have utilized a temporal sliding window approach. In the experiments, we set the window size to twenty five frames and the sliding step size to eight frames.

B. *Person Detection and Tracking*

For the detection of persons, we have used HOG descriptors [11] and a pre-trained, publicly available full-body person detector [30] which yields a sparse set of detections in time, i.e. there are a lot of misdetections. False negatives can be solved using tracking approaches, which are anyway needed to provide time coherence to detections, so that we can reconstruct the trajectory of objects.

For the tracking, we have implemented a Rao-Blackwellized Data Association Particle Filter (RB-DAPF) [12]. This type of filter has been proven to provide good multiple object tracking results even in the presence of sparse detections as the ones we have in these sequences, and can be tuned to handle occlusions. The Rao-Blackwellization can be understood as splitting the problem into linear/Gaussian and non-linear/non-Gaussian parts. The linear part can be solved with Kalman Filters, while the non-linear one must be solved with approximation methods like particle filters. In our case, the linear part is the position and size of a bounding box that models the persons. The non-linear part refers to the data association that is the process of generating a matrix that links detections (the HOG ones, for instance), with objects or clutter. The association process can be strongly non-linear, thus sampling approaches can be used. In our case we have implemented ancestral sampling [31].

The experimental results have shown that this approach is able to detect and track persons whose full body is clearly seen in the scene, up to four or five simultaneous persons. When more than five persons exist, we have found that multiple occlusions happen and the full-body detector does not provide good detection results.

The control of input/output of new persons is handled thanks to the use of the data association filter that classifies detections according to the existing objects, removes objects that have no detection for a too long period of time, and creates new objects when detections not associated to previous objects appear repeatedly.

V. Video Quality Measurements and Performance Evaluation of Detectors

In this section, experimental results about the accuracy of the event detectors in relation to the input video quality are presented. Firstly, video quality metrics (i.e., PSNR and SSIM) are demonstrated for videos after transcoding. Afterwards, the effect of video transformations to the performance of detectors is investigated. Video quality metrics for experiments with different frame rates are not demonstrated, since both adopted quality metrics are frame-based and frame synchronization cannot be achieved.

The original videos have been selected from TREC Video Retrieval Evaluation (TRECVID) collection [32]. The TRECVID database is sponsored by the National Institute of Standards and Technology (NIST), with additional support from other U.S. government agencies. The goal of this database is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a benchmark for organizations interested in comparing their results.

The transcoding experiments have been performed with three videos of origin (i.e., LGW_20071101_E1_CAM1.mpeg, LGW_20071101_E1_CAM3.mpeg, LGW_20071101_E1_-CAM5.mpeg) selected from TRECVID videos collection. All videos have the same initial format (MPEG-2) and the same encoding details. The full details of the input videos are included in Table II. The FFmpeg command line application was used for the video transcoding operation [33].

A. CCTV video quality measurements.

Trying to avoid unpredictable spikes in bit rate, constant bit rate encoding is used. The flat bit rate allows a smoother playback with the drawback of a larger file. Figures 1, 3, and 5 demonstrate the PSNR metric (calculated from (1) and (2)) for the three videos of reference, while Figures 2, 4, 6 present the SSIM metric (computed though (3)) of the aforementioned videos. As it is expected, the videos with smaller bit rates have a downgraded video quality. In all cases, the curves of different bit rates follow the same trend over time in the three videos.

$ mediainfo LGW_20071101_E1_CAM1.mpeg	
Complete name:	LGW_20071101_E1_CAM1.mpeg
Format:	MPEG-PS
File size:	5.34 GiB
Duration:	2h 4mn
Overall bit rate mode:	Variable
Overall bit rate:	6124 Kbps
VideoID:	224 (0xE0)
Format:	MPEG Video
Format version:	Version 2
Format profile:	Main@Main
Format settings, BVOP:	Yes
Format settings, Matrix:	Default
Format settings, GOP:	M=3, N=12
Bit rate mode:	Variable
Bit rate:	6002 kbps
Maximum bit rate:	9000 kbps
Width:	720 pixels
Height:	576 pixels
Display aspect ratio:	4:3
Frame rate:	25 fps
Standard:	PAL
Color space:	YUV
Chroma subsampling:	4:2:0
Bit depth:	8 bit
Scan type:	Progressive
Compression mode:	Lossy
Bits/(Pixel*Frame):	0.579
Stream size:	5.24 GiB (98%)

TABLE II. Video file input details.

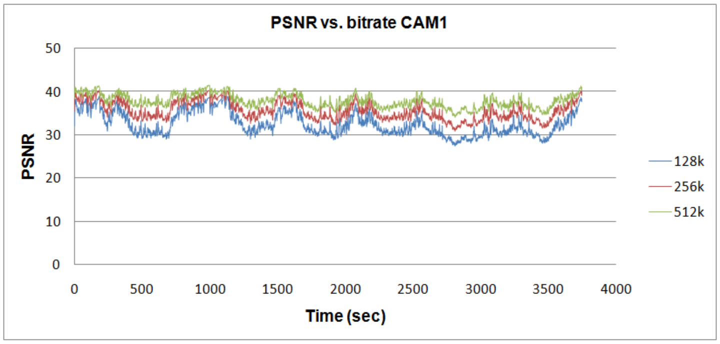

Fig. 1. CAM1: PSNR measure over time as a function of bit rate.

For saving on bit rate, videos can be scaled down to a smaller size by simply lowering the video resolution. Note that it is desirable to maintain the image aspect ratio when resizing. The three original videos have been captures in 4:3 ratio, and changing this ratio can lead to a squishing or stretching effect that is non observable. Figures 7, 9, and 11 present the PSNR metric for the three videos of reference, while Figures 8, 10, 12 present the SSIM metric of the aforementioned videos. The videos with smaller resolution have a downgraded video quality as it is anticipated. The curves of different resolutions follow the same trend over time in the three videos.

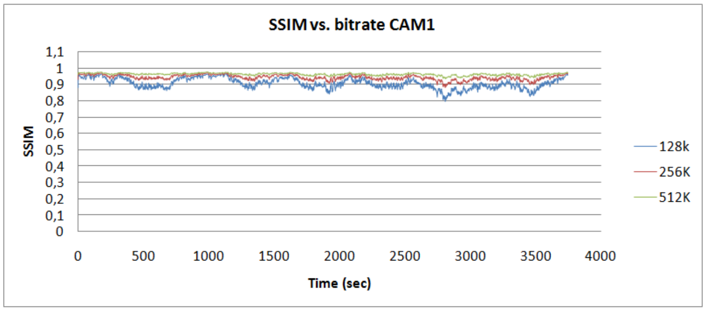

Fig. 2. CAM1: SSIM measure over time as a function of bit rate.

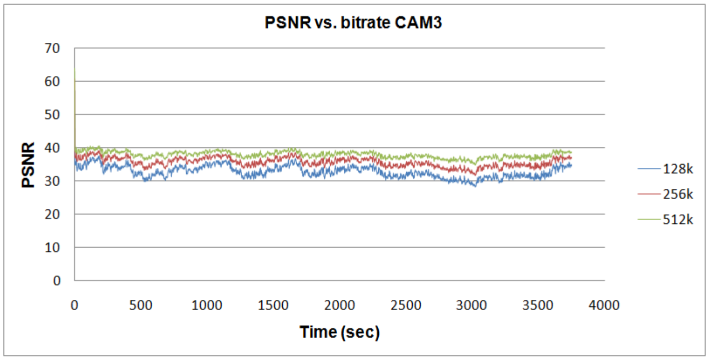

Fig. 3. CAM3: PSNR measure over time as a function of bit rate.

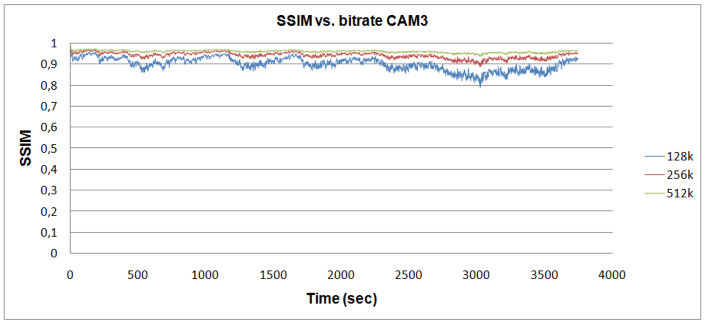

Fig. 4. CAM3: SSIM measure over time as a function of bit rate.

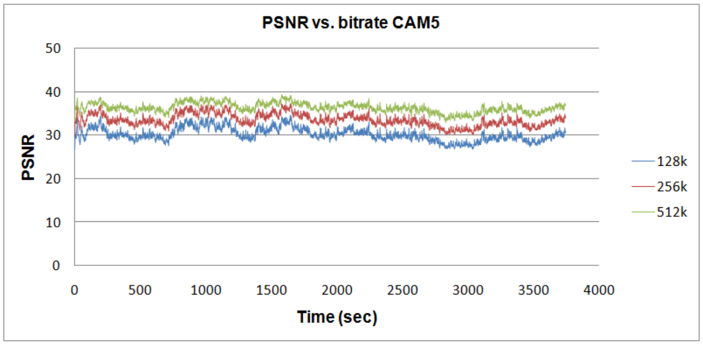

Fig. 5. CAM5: PSNR measure over time as a function of bit rate.

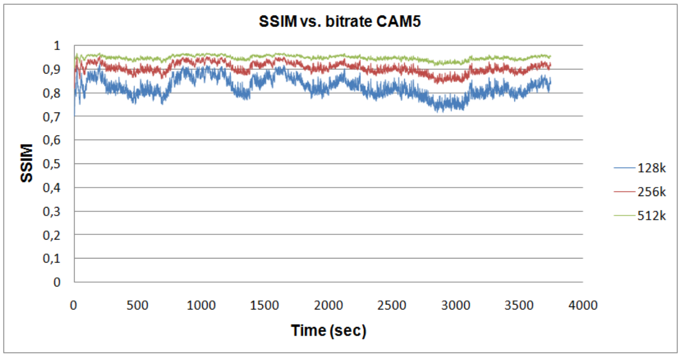

Fig. 6. CAM5: SSIM measure over time as a function of bit rate.

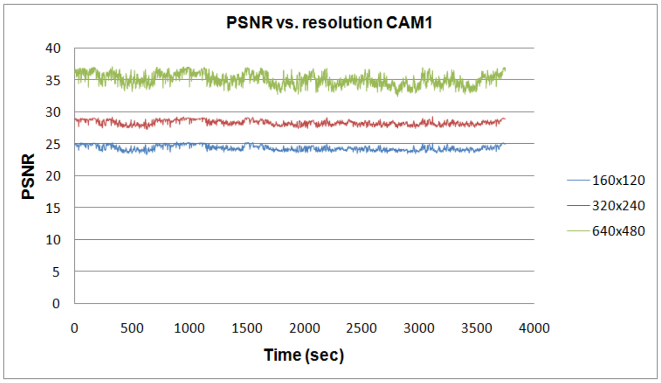

Fig. 7. CAM1: PSNR measure over time for different resolutions.

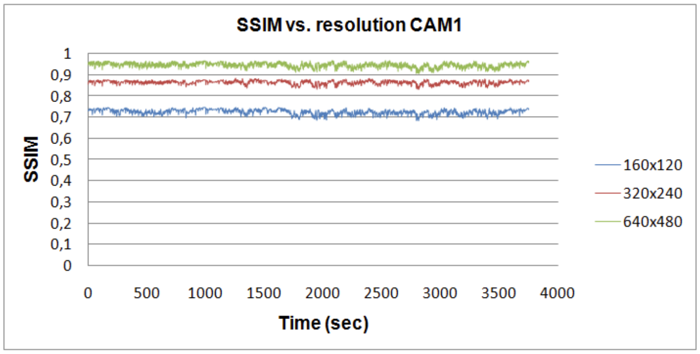

Fig. 8. CAM1: SSIM measure over time for different resolutions.

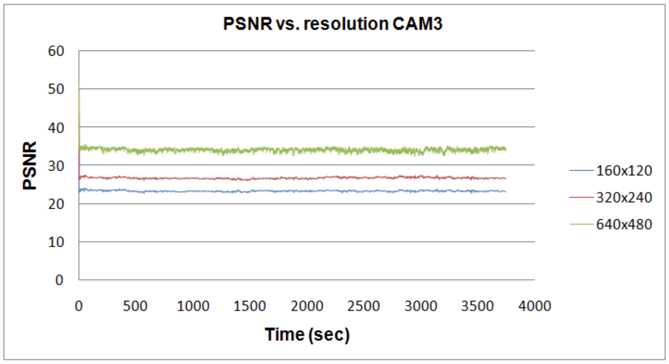

Fig. 9. CAM3: PSNR measure over time for different resolutions.

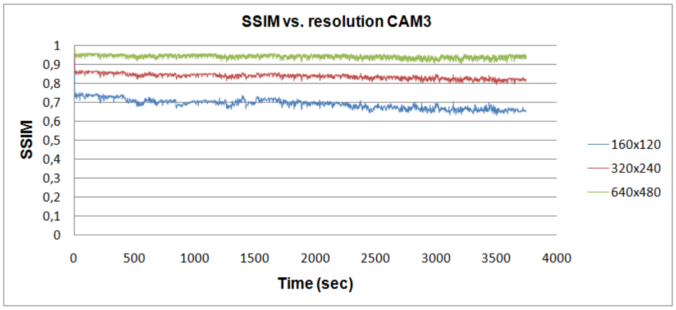

Fig. 10. CAM3: SSIM measure over time for different resolutions.

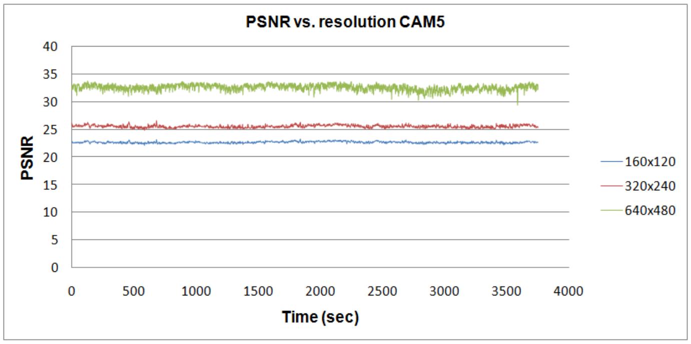

Fig. 11. CAM5: PSNR measure over time for different resolutions.

Regarding the video quality measurements, the three videos demonstrated comparable behaviour in both PSNR and SSIM metrics. Therefore, the same performance is anticipated for a larger collection of CCTV security videos.

B. Evaluation of event detection performance with respect to the degradation in video quality.

In order to check how the video quality affects the CCTV video analysis tasks, we performed person and motion detection tests on the videos obtained using the aforementioned encoder parameters.

At first, we tried to detect persons in the three selected TRECVid videos and then we compared the original results with those obtained from the set of transformed sequences. We have used a subjective metric to define False Positive (FP), False Negative (FN) and True Positive (TP) events, considering that detections from HOG-SVM will be followed by tracking algorithms based on tracklets [30]. For this reason, we have defined these events as inter-frame rates, i.e., TP: a sufficient number of detections of a person along its path on the sequence ($>$ 50% of frames in the sequence); FN: not enough detections along its path ($<$ 50%); FP: a persistent (more than 3 consecutive frames) set of false detections in the same region.

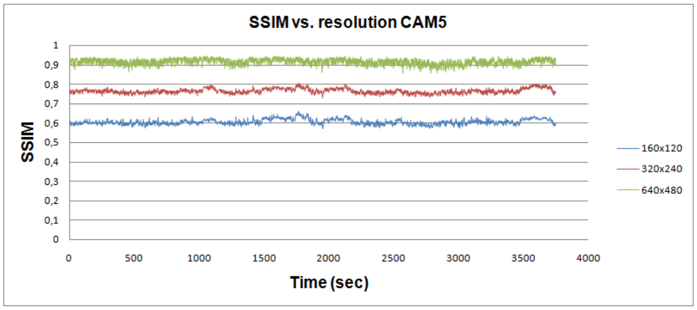

Fig. 12. CAM5: SSIM measure over time for different resolutions.

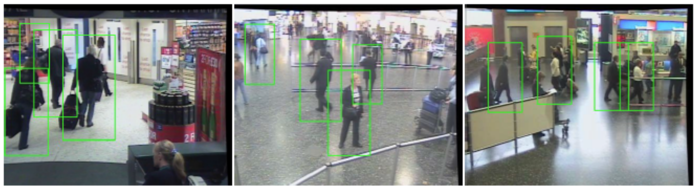

Fig. 13. Example frames of the three cameras used with detections of persons using the HOG-SVM detector [11].

Tables III-V summarize the obtained values of Recall $R \triangleq TP/(TP + FN)$ and Precision $P \triangleq TP/(TP + FP)$ for the different videos, considering the reduction of bit rate, resolution and frame rate. High recall means that an algorithm returned most of the relevant results, while high precision means that an algorithm returned substantially more relevant results than irrelevant [34].

LGW_20071101_E1_CAM1.mpeg

		True Positives	False Positives	False Negatives	Recall	Precision
bit rate	128k	39	17	9	0.81	0.7
	256k	39	13	9	0.81	0.75
	512k	40	12	8	0.83	0.77
frame per sec	5fps	38	4	10	0.79	0.9
	10fps	38	7	10	0.79	0.84
	15fps	38	9	10	0.79	0.81
	20fps	39	10	9	0.81	0.8
	25fps	40	12	8	0.83	0.77
resolution	160x120	40	26	8	0.83	0.61
	320x240	40	20	8	0.83	0.67
	640x480	40	14	8	0.83	0.74

TABLE III. PERSON DETECTION RESULTS FROM CAM1.

LGW_20071101_E1_CAM2.mpeg

		True Positives	False Positives	False Negatives	Recall	Precision
bit rate	128k	37	69	112	0.25	0.35
	256k	43	64	106	0.29	0.4
	512k	48	60	101	0.32	0.44
frame per sec	5fps	34	4	115	0.23	0.89
	10fps	37	7	112	0.25	0.84
	15fps	40	9	109	0.27	0.82
	20fps	44	10	105	0.3	0.81
	25fps	49	12	100	0.33	0.8
resolution	160x120	40	69	109	0.27	0.37
	320x240	43	66	106	0.29	0.39
	640x480	47	59	102	0.32	0.44

TABLE IV. PERSON DETECTION RESULTS FROM CAM2.

The analysis of the results reveals that the detection of objects is affected negatively by the reduction of bit rate (more blockiness effect) and resolution (less information), in both recall and precision. However, the reduction of frame rate increases unexpectedly the precision values; the lower frame

LGW_20071101_E1_CAM3.mpeg						
		True Positives	**False Positives**	**False Negatives**	**Recall**	**Precision**
bit rate	128k	86	39	36	0.7	0.69
	256k	88	37	34	0.72	0.7
	512k	88	34	34	0.72	0.72
frame per sec	5fps	63	8	59	0.52	0.89
	10fps	66	12	56	0.54	0.85
	15fps	70	14	52	0.57	0.83
	20fps	80	25	42	0.66	0.76
	25fps	90	28	32	0.74	0.76
resolution	160x120	66	37	56	0.54	0.64
	320x240	69	36	53	0.57	0.66
	640x480	81	31	41	0.66	0.72

TABLE V. PERSON DETECTION RESULTS FROM CAM3.

rates affect negatively the True Positives and False Negatives, but also decrease False Positives and consequently increase the precision values.

In the experiments using different frame rates, we have observed that the apparently worsen video quality gives higher precision results for the case of human detection. This can be explained by the fact that the basic principle in applying detection by classifiers consists of the use of classifiers trained upon sets that contain thousands of examples of human full-bodies. Each classifier obtains a descriptor that fits the best to the largest subset of the training set. For this reason the descriptors tend to simplify the details of particular figures (full-bodies) and characterize their coarse visual features or appearance. One can think about this process as an averaging, although classifiers can be much more complex than that. With reduced image quality, the details of the video are lost but the image appears smoother than the original (at least in the case of the applied H.264 codec).

As regarding the motion detection experiments, our classifier has been trained to identify 'Pointing Events'. The video entitled LGW_20071101_E1_CAM1 contains six segments with pointing events. In this case, precision is the percentage of the six samples that were correctly identified. The mean confidence is defined as the mean of the normalized (min/max) confidence value and it is a measure of the confidence of classification decision. Note that the classifier has been trained on samples with the same bit rate, frame rate, and resolution as the original video.

LGW_20071101_E1_CAM1.mpeg				
		Correct	**Precision**	**Mean Confidence**
bit rate	128k	3	0.5	0.71
	256k	5	0.83	0.8
	512k	4	0.66	0.8
frame per sec	5fps	1	0.16	0.62
	10fps	0	0	0.63
	15fps	2	0.33	0.69
	20fps	3	0.5	0.73
	25fps	2	0.33	0.76
resolution	160x120	0	0	0.16
	320x240	0	0	0.48
	640x480	2	0.33	0.7

TABLE VI. POINTING DETECTION RESULTS FROM CAM1.

The poor performance of pointing detector in videos of higher quality is not surprising and it can be explained as before. Therefore, PSNR and SSIM (and more general the legacy video quality metrics) are not suitable to demonstrate the degree (positive or negative) that event detectors are affected by the compression. Video quality metrics directly targeting to computer vision applications would be required in this case.

VI. CONCLUSIONS

The performance of human and motion detection algorithms, like the ones analyzed in this paper, is highly affected (either positive or negative) by reductions of bit rate, frame rate and resolution. However, the widely used video quality metrics PSNR and SSIM cannot provide any information or intuition about the change of the precision metrics.

In this work, we estimate the critical video quality for person and motion detection using TRECVID open data set and two common event detectors. In a future work, emphasis will be given in defining novel video quality metrics that are appropriate for CCTV video analysis tasks and computer vision applications, in general.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 285621, project titled SAVASA.

REFERENCES

[1] "Chicago's video surveillance cameras: A pervasive and unregulated threat to our privacy," ACLU of Illinois, Tech. Rep., February 2011.

[2] "FP7 SEC SAVASA Project: Standards-based Approach to Video Archive Search and Analysis, http://www.savasa.eu."

[3] C. Iso/ie, "ISO 13818-2 MPEG2," 1995.

[4] Recommendation ITU-T H.264 — International Standard ISO/IEC 14496-10, April 2013.

[5] Y. Wu, L. Jiao, G. Wu, E. Chang, and Y. Wang, in *in Proc. of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS03), IEEE, Los Alamitos, CA*.

[6] A. M. Eskicioglu and P. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, 1995.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] C. Harris and M. Stephens, "A combined corner and edge detector." in *Proc. Alvey Vision Conference*, 1988, pp. 147–152.

[9] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision." in *Proc. International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[10] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features." in *Carnegie Mellon University Technical Report CMU-CS-91-132*, 1991.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection." in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[12] A. Doucet, N. J. Gordon, and V. Krishnamurthy, "Particle filters for state estimation of jump markov linear systems." *IEEE Trans. on Signal Processing*, vol. 49, no. 3, pp. 613–624, 2001.

[13] S. Little, I. Jargalsaikhan, C. Direkoglu, N. E. O. Connor, K. Clawson, H. Li, M. Nieto, A. Rodriguez, P. Sanchez, K. Paniza, A. M. Llorens, R. Gimenez, R. S. de la Camara, and A. Mereu, "SAVASA Project@ TRECVID 2012: Interactive Surveillance Event Detection, NIST TRECVID Workshop 2012."

[14] P. Rouse and S. Hemami, "Analyzing the role of visual structure in the recognition of natural image content with multi-scale ssim." in *Proc. of SPIE Human Vision and Electronic Imaging XIII Conference, San Jose, CA, USA*, vol. 6806, 2008.

[15] P. Korshunov and W. T. Ooi, "Video quality for face detection, recognition, and tracking," *ACM Trans. on Multimedia Computing, Communications and Applications (TOMCCAP)*, vol. 7, no. 3, p. 14, 2011.

[16] M. Leszczuk and J. Dumke, "The Quality Assessment for Recognition Tasks (QART), VQEG," http://www.its.bldrdoc.gov/vqeg/project-pages/qart/qart.aspx., July 2012.

[17] ITU-R, "Methology for the subjective assessment of the quality of television pictures. Recommendation BT.500-7 (Revised)." 1996.

[18] T. Contin and L. Alpert, "DSCQE Experiment for the Evaluation of the MPEG-4 VM on Error Robustness Functionality." ISO/IEC JTC1/SC29/WG11, MPEG 97/M1604., 1997.

[19] F. Pereira and T. Alpert, "MPEG-4 Video Subjective Test Procedures and Results." *IEEE Trans. on Circuits and Systems for Video Technology.*, vol. 7, no. 1, pp. 32–51, 1997.

[20] "ITU-R Recommendation BT.500-11 – Methodology for the subjective assessment of the quality of television pictures," 2002.

[21] "ITU-T Recommendation P.910 – Subjective video quality assessment methods for multimedia applications," 1999.

[22] M. Ghanbari and K. T. Tan, "A Multi-Metric Objective Picture Quality Measurements Model for MPEG video." *IEEE Trans. on Circuits and Systems for Video Technology.*, vol. 10, no. 7, pp. 1208–1213, 2000.

[23] M. H. Wolf and S. Pinson, "Spatial – Temporal Distortion Metrics for in-service Quality Monitoring of any Digital Video System." in *Proc. SPIE International Symposium on Voice, Video, and Data Communications, Boston.*, 1999, pp. 11–22.

[24] Z. Wang, A. Bovik, H. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 1–14, 2004.

[25] P. Kelly, C. Conaire, C. Kim, and N. E. O. Connor, "Automatic camera selection for activity monitoring in a multi-camera system for tennis." in *Proc. Third ACM/IEEE Int. Conference on Distributed Smart Cameras*, 2009, pp. 1–8.

[26] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3169–3176.

[27] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[28] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition." in *Proc. BMVC 2009-British Machine Vision Conference*, 2009.

[29] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance." in *Proc. European Conference on Computer Vision (ECCV)*, 2006, pp. 428–441.

[30] A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006, pp. 666–673.

[31] C. R. del Blanco, F. Jaureguizar, and N. Garcia, "An advanced Bayesian model for the visual tracking of multiple interacting objects." *EURASIP Journal on Advances in Signal Processing*, vol. 130, 2011.

[32] TREC Video Retrieval Evaluation (TRECVID), http://www-nlpir.nist.gov/projects/trecvid.

[33] "FFmpeg project. http://www.ffmpeg.org."

[34] "Precision and recall, http://en.wikipedia.org/wiki/precision_and_recall."