# 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
{mkafetz, cxilouris, akis.kourtis}@iit.demokritos.gr

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

Iveel Jargalsaikhan and Suzanne Little
CLARITY Centre
for Sensor Web Technology
Dublin City University, Ireland
iveel.jargalsaikhan2@mail.dcu.ie
suzanne.little@dcu.ie

***Abstract*—The process of transcoding videos apart from being computationally intensive, can also be a rather complex procedure. The complexity refers to the choice of appropriate parameters for the transcoding engine, with the aim of decreasing video sizes, transcoding times and network bandwidth without degrading video quality beyond some threshold that event detectors lose their accuracy. This 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 and resolution, while motion detection is more sensitive to frame rate.**

***Keywords*-CCTV; transcoding; event detection; video quality.**

## I. Introduction

The existing Closed-Circuit Television (CCTV) infrastructures and surveillance video systems are not 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. The SAVASA project [1] 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. At the core of the search interface is the application of algorithms for person/object detection and tracking.

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, watermarking and storage. In this paper, conversions between MPEG-2 coding standard to H.264/MPEG-4 Advanced Video Coding (AVC) standard are performed.

The key disadvantage of transcoding is that more frequently it is a lossy process, introducing image artifacts 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 [2], they do not have high-quality video requirements. In future surveillance systems, videos will be mainly transmitted for processing by automated video analysis algorithms, with the minimum acceptable quality to increase the scalability.

Nevertheless, 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) [3] and Frame Rate Structural SIMularity (SSIM) [4] 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 [5] and tracked using the Kanade-Lucas-Tomasi (KLT) algorithm [6], [7]. 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 [8] and tracked via Rao-Blackwellized Data Association Particle Filter [9].

The lowest video quality allowing humans to perform recognition of natural image contents is studied in [10]. From computer vision perspective, the most relevant work to ours is [11] 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.

## II. Transcoding Parameters

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

*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. This flexibility allows smaller overall file sizes without serious compromises in the quality of the video.

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 order to have a broad picture of bit rates in CCTV systems, note that one camera might produce between 100 kbps and 2 Mbps.

*Video Resolution*: Resolution is a measurement of the number of pixels in a frame. As more pixels exist in the frame, the image gets sharper and more detailed. Typically, the resolution is expressed as frame length times height (both measured in pixels). Common resolutions for CCTV IP cameras are the CIF ($640 \times 480$), the 4CIF ($704 \times 480$) and the D1($720 \times 480$). The resolution $1280 \times 720$ is the minimum that is called High Definition (HD). There is also $1920 \times 1080$ resolution, which is sometimes referred to as Full HD.

*Frame Rate*: This parameter specifies the number of frames that are generated 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 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 realise 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 [12]. Since all these works are in a very early stage, we review here only well established video quality metrics, categorised into two broad classes: the subjective and the objective ones.

*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) [13]–[15]. However, the MOS method is usually inconvenient due to the fact that the preparation and execution of subjective tests is costly and time consuming.

*Objective quality*: 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 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 characterised as Full Reference Methods (see, e.g., [16], [17]). 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 the development of several objective video quality models, PSNR continues to be the most popular evaluation of the quality difference among videos, i.e., $\text{PSNR} \triangleq 10\log_{10}(L^2/\text{MSE})$, where $L$ denotes the dynamic range of pixel value. The MSE is defined by $\text{MSE} \triangleq \sum_{i=1}^{N}(x_i - y_i)^2/N$, where $N$ denotes the number of pixels, and $x_i$, $y_i$ the $i^{\text{th}}$ pixel in original, distorted frame, respectively.

Frame Rate Structural SIMularity (SSIM) is a 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 specialised in extracting the errors. If $x$ and $y$ are two video frames, $\text{SSIM}(x,y) \triangleq \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$, where $\mu_x$, $\mu_y$ are the mean value of $x$ and $y$, $\sigma_x$, $\sigma_y$, $\sigma_{xy}$ are the variances of $x$, $y$ and the covariance of $x$ and $y$, respectively. The constants $C_1$ and $C_2$ are defined as $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$, and $K_1 = 0.01$ and $K_2 = 0.03$, respectively [18]. The SSIM gives more reliable result than PSNR. However, the lower computational complexity of PSNR makes it ideal to apply in real-time applications.

## IV. Person and Motion Detection

This section outlines the two classifiers used to identify video segments that include one of the two following events: a) Pointing (as used in the TRECVid surveillance event detection task [19]) and b) Person-Walks.

*Pointing Detection Using Motion Trajectory*: To represent motion, we have used salience points for capturing the motion trajectory. This low-level feature is described by four different descriptors. Firstly, in order to facilitate motion trajectory extraction, a background subtraction algorithm [20] 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 [5] and are tracked over video sequences using Kanade-Lucas-Tomasi (KLT) algorithm [6], [7]. 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 have adopted the approach in [21] to describe the trajectory features. For each trajectory, we have calculated four descriptors to capture the different aspects of motion trajectory. Among the existing descriptors, HOG/HOF [22] has shown to give excellent results on a variety of datasets [23]. Therefore, HOG/HOF is computed along our trajectories. HOG (Histogram of Oriented Gradient) [8] captures the local appearance around the trajectories, whereas HOF (Histogram of Optical Flow) captures the local motion [21]. Additionally, MBH (Motion Boundary Histogram) [24] and TD (Trajectory Descriptor) [21] 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 was set to k=4000, which has shown empirically to give good results in [22]. 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.

*Person Detection and Tracking*: For the detection of persons, we have used HOG descriptors [8] and a pre-trained, publicly available full-body person detector [25] which yields a sparce 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) [9]. 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 [26].

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 Event 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 are demonstrated for videos after transcoding. Afterward, the effect of video transformations to the performance of detectors is investigated[1].

The original videos have been selected from TREC Video Retrieval Evaluation (TRECVID) collection [19]. The transcoding experiments have been performed with three scenes of two minutes duration, each one taken from the beginning of three original videos from three cameras. All videos have the same initial format (MPEG-2) and the same encoding details. The bit rate of input videos is variable with mean value 6002 kbps, the frame rate equals to 25 fps and the resolution is $720 \times 576$. The FFmpeg application was used for the video transcoding operation [27], while the MSU Video Quality Measurement Tool [28] has been applied for measuring video quality.

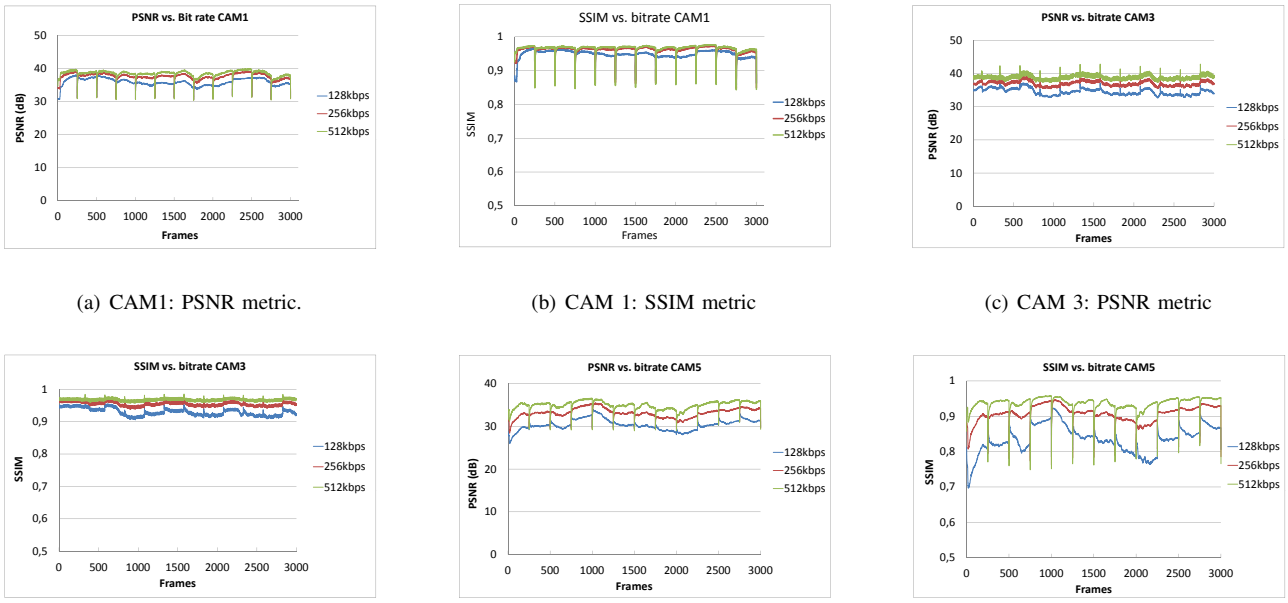
*CCTV video quality measurements.*: Trying to avoid unpredictable fluctuations in bit rate, constant bit rate encoding has been used. Subfigures 1(a), 1(c), and 1(e) demonstrate the PSNR metric for the three videos of reference, while Subfigures 1(b), 1(d), 1(f) present the SSIM metric (computed though (III)) of the aforementioned videos. Each video has been transcoded in three different bit rates: 128 kbps, 256 kbps, and 512 kbps.

As expected, the videos with smaller bit rates have downgraded video quality. The curves of different bit rates follow the same trend over time in the three videos[2].

For saving on bit rate, videos have been scaled down to a smaller size by simply lowering the video screen resolution. The three original videos have been captured in 4:3 ratio. In order to maintain the original image aspect ratio in the resizing, the videos have been transcoded to $160 \times 120$, $320 \times 240$, and $640 \times 280$ screen resolutions. Subfigures 2(a), 2(c), and 2(e) present the PSNR metric for the three videos of different resolution, while Subfigures 2(b), 2(d), and 2(f)

[1]Video quality metrics for experiments with different frame rates are not demonstrated, since both adopted quality metrics are frame-based and frame synchronisation cannot be achieved.

[2]The periodic spikes that appear every 250 frames (default value for I-frame interval in MPEG-2 format) are caused by the fact that I-frames are of higher video quality than P and B frames.




(a) CAM1: PSNR metric. (b) CAM 1: SSIM metric (c) CAM 3: PSNR metric

(d) CAM 3: SSIM metric (e) CAM 5: PSNR metric (f) CAM 5: SSIM metric

Fig. 1. Video quality measurements for each frame as a function of bit rate.

demonstrate 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.

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.

*Evaluation of event detection performance with respect to the degradation in video quality.*: In order to investigate how the video quality affects the CCTV video analysis tasks, we have performed person and motion detection tests on the videos obtained using the aforementioned encoder parameters.

At first, we have tried to detect persons, and in the following we have 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 the detections from HOG-SVM will be followed by tracking algorithms based on tracklets [25]. Thus, 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.

Tables I-III summarise 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.

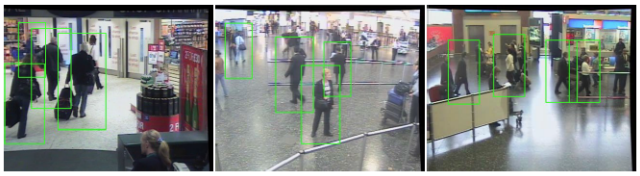
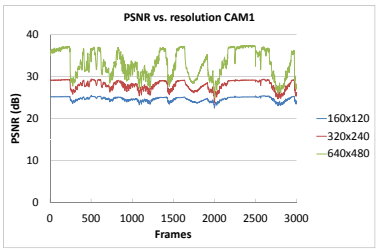


Fig. 3. Example frames of the three cameras used with detections of persons using the HOG-SVM detector [8].

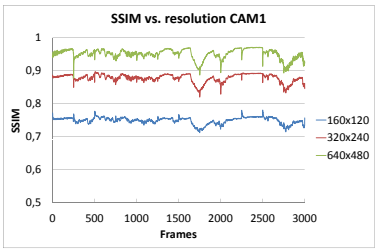
| LGW_20071101_E1_CAM1.mpeg | | | | | |
|---|---|---|---|---|---|
| | **True Positives** | **False Positives** | **False Negatives** | **Recall** | **Precision** |
| original | 40 | 11 | 8 | 0.83 | 0.78 |
| 128kbps | 39 | 17 | 9 | 0.81 | 0.7 |
| 256kbps | 39 | 13 | 9 | 0.81 | 0.75 |
| 512kbps | 40 | 12 | 8 | 0.83 | 0.77 |
| 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 |
| $160 \times 120$ | 40 | 26 | 8 | 0.83 | 0.61 |
| $320 \times 240$ | 40 | 20 | 8 | 0.83 | 0.67 |
| $640 \times 480$ | 40 | 14 | 8 | 0.83 | 0.74 |

TABLE I
PERSON DETECTION RESULTS FROM CAM1.

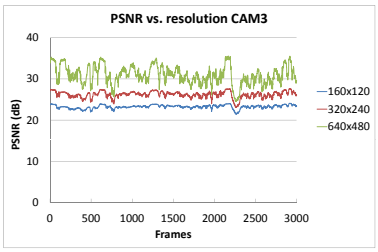
The analysis of the results reveals that the detection of persons is affected negatively by the reduction of bit rate (more blockiness effect) and resolution (less information), in



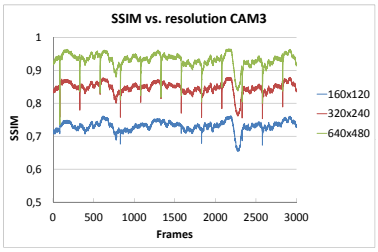

(a) CAM1: PSNR metric.



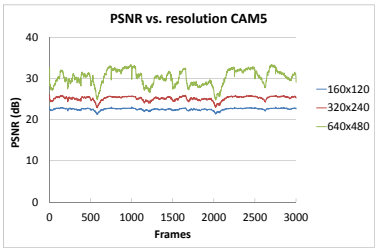

(b) CAM 1: SSIM metric



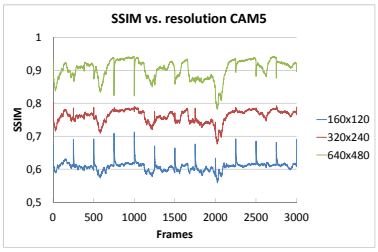

(c) CAM 3: PSNR metric




(d) CAM 3: SSIM metric




(e) CAM 5: PSNR metric




(f) CAM 5: SSIM metric

Fig. 2. Video quality measurements for each frame for different resolutions.

| LGW_20071101_E1_CAM3.mpeg | | | | | |
|---|---|---|---|---|---|
| | **True Positives** | **False Positives** | **False Negatives** | **Recall** | **Precision** |
| original | 49 | 58 | 100 | 0.33 | 0.46 |
| 128kbps | 37 | 69 | 112 | 0.25 | 0.35 |
| 256kbps | 43 | 64 | 106 | 0.29 | 0.4 |
| 512kbps | 48 | 60 | 101 | 0.32 | 0.44 |
| 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 | 0.30 | 0.81 |
| 25fps | 49 | 12 | 100 | 0.33 | 0.80 |
| $160 \times 120$ | 40 | 69 | 109 | 0.27 | 0.37 |
| $320 \times 240$ | 43 | 66 | 106 | 0.29 | 0.39 |
| $640 \times 480$ | 47 | 59 | 102 | 0.32 | 0.44 |

TABLE II
PERSON DETECTION RESULTS FROM CAM3.

| LGW_20071101_E1_CAM5.mpeg | | | | | |
|---|---|---|---|---|---|
| | **True Positives** | **False Positives** | **False Negatives** | **Recall** | **Precision** |
| original | 90 | 28 | 32 | 0.74 | 0.76 |
| 128kbps | 86 | 39 | 36 | 0.70 | 0.69 |
| 256kbps | 88 | 37 | 34 | 0.72 | 0.7 |
| 512kbps | 88 | 34 | 34 | 0.72 | 0.72 |
| 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 |
| $160 \times 120$ | 66 | 37 | 56 | 0.54 | 0.64 |
| $320 \times 240$ | 69 | 36 | 53 | 0.57 | 0.66 |
| $640 \times 480$ | 81 | 31 | 41 | 0.66 | 0.72 |

TABLE III
PERSON DETECTION RESULTS FROM CAM5.

both recall and precision. In the experiments using different frame rates, we have observed that the apparently worse video quality gives higher precision results for the case of human body detection. In other words, the reduction of frame rate increases unexpectedly the precision values; the lower frame rates affect negatively the True Positives and False Negatives, but also decrease drastically the False Positives. Consequently, the lower frame rate increase the precision values. This can be justified by the fact that it is more difficult to occur three consecutive false detections (in order to trigger one False Positive event) when the frame rate is decreased. This behaviour is not observed for the recall metric, since False Positives are not taken into account for its computation.The increase of precision due to the lower frame rates cannot be captured by the video quality metrics, revealing that the existing video quality metrics are not suitable for automated systems. Video quality metrics directly targeting to computer vision applications would be required in this case.

For the motion detection experiments, our classifier has been trained to identify 'Pointing Events'. The video subset from LGW_20071101_E1_CAM1 contains six segments with pointing events. In this case, the precision is the percentage of samples that were correctly identified. The mean confidence is defined as the mean of the normalised (min/max) confidence interval value, and it can be considered as an indicator of the confidence of the classification decision. Note that the

classifier has been trained on samples with the same bit rate, frame rate, and resolution as the original video, and it can classify correctly each of the original video segments.

Table IV presents the results of classifying 'Pointing' events. Due to the small sample size, a miss in detecting an event affects severely the precision values. For instance, the classifier has very poor accuracy with the resolution of $160 \times 120$. However, it also has low mean confidence, so a threshold could be applied to discard such values.

In general, reducing the frame rate is poor choice, since the trajectory-based approach does not work efficiently if there are missing frames. These preliminary results do not show strong correlation with the video quality metrics but do indicate that an acceptable trade-off between performance and video size may be possible by reducing the bit rate. In all cases, the higher bit rate, frame rate and resolution is beneficial for increasing the confidence of precision values.

| LGW_20071101_E1_CAM1.mpeg | | | | |
|---|---|---|---|---|
| | | **Correct** | **Precision** | **Mean Confidence** |
| bit rate | 128kbps | 3 | 0.50 | 0.71 |
| | 256kbps | 5 | 0.83 | 0.80 |
| | 512kbps | 4 | 0.66 | 0.80 |
| frame rate | 5fps | 1 | 0.16 | 0.62 |
| | 10fps | 0 | 0 | 0.63 |
| | 15fps | 2 | 0.33 | 0.69 |
| | 20fps | 3 | 0.50 | 0.73 |
| | 25fps | 2 | 0.33 | 0.76 |
| resolution | $160 \times 120$ | 0 | 0 | 0.16 |
| | $320 \times 240$ | 0 | 0 | 0.48 |
| | $640 \times 480$ | 2 | 0.33 | 0.70 |

TABLE IV
POINTING DETECTION RESULTS FROM CAM1.

## VI. CONCLUSIONS AND FUTURE WORK

The performance of human body and motion detection algorithms, like the ones analysed in this paper, is highly affected 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 a subset from the TRECVID 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.

## ACKNOWLEDGMENT

This work has received funding from the EU FP7 research project SAVASA (285621).

## REFERENCES

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

[2] Y. Wu, L. Jiao, G. Wu, E. Chang, and Y. Wang, "Invariant feature extraction and biased statistical inference for video surveillance." in *Proc. of the IEEE Int. Conference on Advanced Video and Signal-Based Surveillance (AVSS03), IEEE, Los Alamitos, CA*, 2003, pp. 284–289.

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

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid." in *Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR' 06*, 2006, pp. 321–330.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] "MSU video quality measurement tool, http://compression.ru-/video/quality_measure/video_measurement_tool_en.html."