2012 IEEE International Conference on Multimedia and Expo

ACCURACY AND POWER CONSUMPTION TRADEOFFS IN VIDEO RATE ADAPTATION FOR COMPUTER VISION APPLICATIONS

Yousef O. Sharrab and Nabil J. Sarhan

Electrical & Computer Engineering Department, Wayne State Media Research Lab
Wayne State University, Detroit, MI 48202
{yousef.sharrab,nabil}@wayne.edu

ABSTRACT

This paper analyzes and compares the rate-accuracy and rate-energy characteristics of various video rate adaptation techniques in computer vision applications. The analyzed rate adaptation techniques include spatial, spatial with upscaling, temporal, and Signal-to-Noise Ratio (SNR). We experiment with standard video sequences as well as 300 security, surveillance, news, and speech videos. These videos total 19.15 hours of recording time. We consider both MPEG-4 and H.264 compression standards.

Index Terms— Automated Video Surveillance, Computer Vision, Power Consumption, Rate-Accuracy, Video Rate Adaptation, Video Streaming.

1. INTRODUCTION

This paper analyzes video rate adaptation techniques in computer vision applications, including Automated Video Surveillance (AVS). As in all other video streaming applications, the video streams in computer vision systems should be adapted to the dynamically changing network conditions. Video rate adaptation has been studied extensively in video streaming in general, but little work has been devoted to computer vision systems in general and AVS systems in particular. For video streaming, a variety of video rate adaptation techniques have been studied [1, 2, 3], with the main approaches being transcoding and scaling the video Signal-to-Noise Ratio (SNR), spatial, and/or temporal parameters.

Most video adaptation techniques considered the video distortion as the primary metric, leading to much literature on rate-distortion characterization and optimization [4] (and references within). In AVS, however, the main objective is enhancing the event/object detection accuracy. The detection accuracy can essentially be thought of as the quality perceived by machines, as opposed to the human perceptual quality or distortion/error/similarity metrics measured by distortion, Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), Structural Similarity Index (SSIM), or other related metrics.

Unfortunately, only few studies have considered the impact of video adaptation on detection accuracy. Study [5] considered the impacts of rate adaptation on accuracy, but only for images and only when SNR adaptation is employed. Study [6] analyzed rate adaptation for only MJPEG videos and did not consider more efficient codecs, such as MPEG-4 and H.264. Moreover, these studies experimented with small datasets and did not consider power consumption.

This paper analyzes and compares various video adaptation techniques in terms of both event/object detection accuracy and power consumption. The analyzed rate adaptation techniques include spatial, temporal, and SNR. We experiment with 9 standard video sequences, and 300 actual security, surveillance, news, and speech videos in a wide variety of resolutions. These videos have a total of 19.15 hours of recording time. We study the videos in both MPEG-4 and H.264 compression standards and assess both the *detection index* and *false positive index*. Finding the probability of false positive for videos was a hard task since it requires human observations of the videos with the imposed markings of detected faces and manual recordings of the results. For spatial adaptation, upscaling the video to original size by the receiver might be beneficial. We experiment with five main upscaling/super-resolution algorithms. Temporal upscaling is also possible but is not considered because it may not be suitable for surveillance applications.

The **main contributions** of this paper can be summarized as follows: (i) analyzing rate adaptation for MPEG-4 and H.264 videos, (ii) considering both the detection accuracy and power consumption, (iii) analyzing the performance of various upscaling algorithms, and (iv) conducting extensive experiments with a large set of actual security, surveillance, news, and speech videos in a wide variety of resolutions, in addition to standard video sequences.

The rest of the paper is organized as follows. Section 2 discusses background information. Section 3 discusses the performance evaluation methodology and Section 4 presents and analyzes the main results. Finally, conclusions are drawn in the last section.

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.77

2. BACKGROUND INFORMATION

As shown in Figure 1, video streaming systems, including automated video surveillance systems, consist of three main phases at the source side: capturing, encoding, and transmission. In video encoding, both intra-prediction and

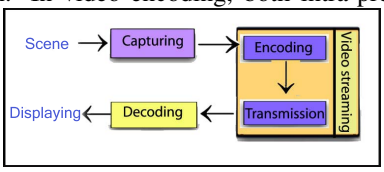


Fig. 1: Block Diagram of a Video Streaming System

inter-prediction are used to reduce the spatial and temporal redundancies in the video, respectively. The first frame of a sequence or a random access point is typically intra-coded. Each block of pixels in an intra-frame is predicted using previously-encoded neighboring blocks. For all remaining frames of a sequence or between random access points, inter-coding is usually used, employing block motion compensation to predict blocks from other previously decoded frames. The residuals of the intra-prediction and inter-prediction are then transformed to the frequency domain using Discrete Cosine Transform (DCT) in MPEG-4 or Integer DCT in H.264. Subsequently, the transform coefficients are quantized, entropy coded, and then transmitted together with any possible motion vectors.

Before the transmission, the video streams may be adapted to fit bandwidth and/or energy constraints [1, 2, 3]. The main approaches for video adaptation include transcoding and scaling the video Signal-to-Noise Ratio (SNR), spatial, and/or temporal parameters. The SNR quality is controlled by changing the quantization parameters of the transform function. The spatial and temporal qualities, however, are controlled by changing the frame size or frame rate, respectively. At the receiver side, the video is decoded. For spatially-adapted videos, an upscaling algorithm may be used to restore the videos to their original sizes. Upscaling (also called super-resolution) is a set of mathematical methods for enhancing the resolution of an image or video by using interpolation. Upscaling algorithms include *Nearest Neighbor*, *Bilinear*, *Bicubic*, *Spline*, and *Lanczos*. The first three algorithms consider the closest pixel, the closest 2×2 pixels, and the closest 4×4 pixels, respectively. Spline and Lanczos consider more surrounding pixels.

3. PERFORMANCE EVALUATION METHODOLOGY

The characteristics of the selected video sequences are summarized in Table 1. These sequences were selected such that each video frame contains exactly one face, thereby simplifying the computation of the detection and false positive indices.

Moreover, we experimented with a dataset, called *WAV-Image* [7], that we collected from YouTube. It has 300 videos of varying quality, among which 100 videos are from actual

Table 1: Characteristics of Selected Video Standard Sequences [The frame rate in each video is 30 fps]

Sequence	Duration (s)	Resolution	# Frames
Carphone	12	QCIF	382
Suzie	5	QCIF	150
Akiyo	10	QCIF	300
Claire	16	QCIF	494
Silent	10	CIF	300
Akiyo	10	CIF	300
Deadline	45.8	CIF	1374
SignIrene	18	CIF	540
vtc1nw	12	4SIF (VGA)	360

Table 2: Characteristics of the Collected 300 Videos [The frame rate in each video is 30 fps]

Description	# Videos	Duration (s)	Resolution	# Frames
Surveillance and Hidden-Camera	100	2857	QVGA	85,710
News and speeches	200	66096	QVGA	1,982,880
Total	300	68953	QVGA	2,068,590

security and surveillance sites. For the sake of comparison, we also collected 200 videos from news, presentation, and speeches of varying quality. These 200 videos generally exhibit better video capturing conditions and quality than the security and surveillance videos. The quality, however, of future AVS systems will improve and may match those of the other set. All these videos were obtained from YouTube. The collection process was mainly based on searching for keywords, such as security, hidden camera, speeches, and news. The videos were chosen such that there are faces in most of the frames. The characteristics of these sequences are summarized in Table 2.

For the video sequences, we used the yuv2avi-p2 program to convert the sequences from raw YUV to AVI format while preserving the original quality. For all videos, we used E.M. Total Video Converter, which is based primarily on FFmpeg, to convert videos to different spatial, temporal, and SNR qualities.

Each video was evaluated at 48 different quality levels, leading to a total of $14,832$ experiments (309×48). The quality levels are obtained by varying the spatial resolution (i.e., frame size), temporal resolution (i.e., frame rate), and SNR resolution (i.e., target bit rate). The frame size was varied to lower settings than the original frame sizes and these settings vary based on the video type. The temporal resolution was varied from 1 to 30 fps (frames per second) for all videos. Furthermore, the SNR setting was varied by changing the target bit rate from 1 Kbps to 240 Kbps for all videos. The videos are encoded in a single pass because of the streaming environment in AVS.

For the spatially-adapted videos, we analyzed the impact of upscaling the videos to their original sizes before

running the computer vision algorithm at the receiver. We experimented with five super-resolution algorithms. We used FFmpeg to upscale the spatially-adapted videos using lossless compression to ensure that no loss in quality happens due to compression.

Subsequently, we conducted the experiments to assess the impact of video adaptation on detection accuracy. As an example, we consider face detection, which is a major algorithm in AVS, and use the Viola-Jones algorithm [8] as implemented in OpenCV library. This algorithm was shown to perform better than other algorithms [9].

In the comparisons between MPEG-4 and H.264, we do not intend to highlight the better coding efficiency of H.264, but rather we aim at comparing the relative rate-accuracy characteristics of these two encoders. Comparing the absolute values of the bitrate between these two encoders is of no essence here because of the varying presets and abilities to achieve the desired bit rate with SNR adaptation.

We use two metrics for the detection accuracy: *average detection accuracy* and *number of detections* (unnormalized detection accuracy). Each one of these metric provides a measure of the detection accuracy but each one is suitable for a different situation. The first metric, also called *detection index*, can be defined as the number of correctly detected faces divided by the total number of faces in all video frames. It is used for the standard sequences, whereas the second metric is used for the other 300 actual videos because the total number of faces in each video is unknown. In addition, we assess the probability of false positive, called (*false positive index*), by observing in slow motion a small subset of the videos with the imposed markings of detected faces and manually recording the results.

The power consumption experiments were conducted on a Dell Inspiron 1525 laptop with a dual-core processor and an external Webcam Pro 9000. The camera feeds a raw video, which is then compressed with FFmpeg using H.264. The power was measured by "Watts Up? Pro ES AC" Graphic Timer Watt meter. The camera was directed to a computer screen playing a specific movie (from the beginning to the end) to ensure that the experiments can be repeated without changes in the video content. Each experiment was repeated four times, and then the overall results were averaged.

4. RESULT PRESENTATION AND ANALYSIS

We study the effectiveness of various adaptation techniques: spatial, spatial with upscaling, temporal, and SNR. In all shown figures, the video parameter that is unchanged by the adaptation technique is set to its largest value. Therefore, the frame size in temporal and SNR adaptations is set to 176×144 in QCIF, 352×288 in CIF, 320×240 in QVGA, and 640×480 in VGA or 4SIF. Similarly, the frame rate in spatial and SNR adaptations is set to 30 fps, and the target bit rate in spatial and temporal adaptations is set to 240 Kbps.

Table 3: Comparing Upscaling Algorithms in % Detection Accuracy [Average results for four QCIF and four CIF sequences. QCIF sequences: Carphone, Suzie, Akiyo, and Claire. CIF sequences: Silent, Akiyo, Deadline, and SignIrene. Codec used: H.264]

Kbps	None	Neighbor	Bilinear	Bicubic	Spline	Lanczos
QCIF Sequences						
9	0.00	3.41	12.12	9.74	8.00	7.60
19	0.00	52.34	53.05	58.44	56.22	58.20
26	0.00	72.38	71.82	73.64	75.62	76.10
30	2.22	79.90	76.65	79.50	80.53	80.37
46	5.46	86.15	80.29	83.38	85.28	85.84
64	23.68	88.37	86.07	89.56	89.95	89.32
CIF Sequences						
34	0.00	5.51	9.59	9.23	9.94	9.62
98	0.00	68.86	71.16	72.13	73.01	72.46
200	0.00	83.96	82.89	85.61	86.87	86.71
240	0.00	89.01	87.65	90.66	89.92	89.43
253	7.03	90.18	88.85	90.11	89.43	90.44
261	29.22	92.12	91.41	90.21	91.34	91.08
	Averages for the above 12 data rows					
	5.63	67.68	67.63	69.35	69.68	69.76

Effectiveness of Upscaling Spatially-Adapted Videos
Let us first discuss the effectiveness of upscaling spatially-adapted videos to their original sizes. Table 3 compares the effectiveness of various upscaling algorithms in terms of the achieved detection accuracy. Only the results for H.264 are shown since MPEG-4 videos exhibit similar characteristics. The results are based on four QCIF and four CIF standard sequences. Each group of four sequences is treated as one long sequence and the overall detection accuracy is reported. Upscaling algorithms can improve the detection accuracy by a factor of 12 to 12.4 on the average. The best performers are Bicubic, Spline, and Lanczos, with Lanczos achieving the highest detection accuracy. These three algorithms vary in the detection accuracy by only 0.41% on the average. Interestingly, Nearest Neighbor performs better than Bilinear, although it has lower complexity. Based on the tradeoff between accuracy and time complexity, Bicubic is the best overall performer, and thus it will be assumed from this point on, unless otherwise indicated.

Comparing Video Rate adaptation Techniques in Detection Accuracy

Results for Standard Sequences: Figure 3 compare the rate-accuracy characteristics under a selected 4SIF sequence. Figures 4 and 5 compare the rate-accuracy characteristics of the four video rate adaptation techniques under four selected CIF sequences, whereas Figures 6 and 7 show the results under four selected QCIF sequences. Each sequence is encoded by both MPEG-4 and H.264 compression standards. These results demonstrate that SNR adaptation and the spatial with upscaling exhibit the best rate-accuracy characteristics. There-

fore, changing the bitrate by varying the quantization parameter or the frame resolution has generally the least negative impact on detection accuracy. SNR performs considerably better than the spatial with upscaling, especially for low bitrates. In "Suzie" and "Akiyo" sequences, the accuracy is nearly constant within the achieved range of the target bitrate in the case of SNR adaptation and spatial with upscaling. Note that in SNR adaptation, the target (i.e., desired) bitrate cannot always be achieved, especially for low bitrates. With spatial adaptation and no upscaling, the detection accuracy is zero below a certain resolution and then increases with the resolution at a high rate until the curve reaches a knee, after which only diminishing returns can be achieved by higher rates.

By upscaling the lower resolutions to the original size by the receiver before running the vision algorithm, the spatial adaptation approaches that of SNR. SNR has the advantage of avoiding the computation complexity and thus the consequent power consumption of running the upscaling algorithm at the receiver.

Since each frame has exactly one face, the detection accuracy changes linearly with the frame rate, but the bitrate is not linear with the frame rate because of the employed compression standards, which exploit temporal correlations among successive frames. Temporal adaptation performs worse than spatial adaptation except for small resolutions. This behavior can be explained as follows. When dropping frames, the faces in these frames will have no chance to be detected. The crossover point between spatial and temporal adaptations varies with the video content and compression scheme.

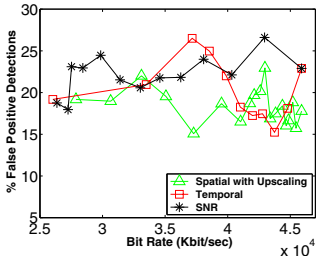
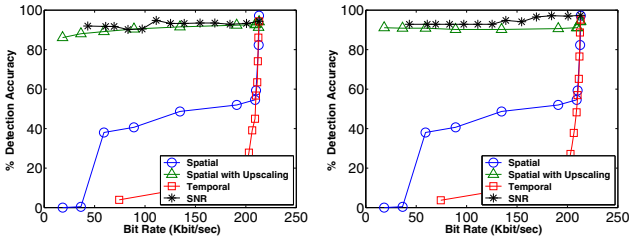


Fig. 2: False Positive Index [Average Results for Four CIF Sequences, MPEG-4 and H.264, Upscaling done with Bicubic Interpolation]

Figure 2 compares the false positive index of various techniques. Spatial with upscaling achieves generally the best in this metric. The variations in the false positive index in the cases of SNR and spatial with upscaling are due to the reduced quality of the background images, especially in Silent and Deadline sequences.

Results for News and Speech Videos: As discussed earlier, for the news and speech videos, we use the number of detected faces as a metric since the number of faces is unknown. These 200 videos are treated as one long video and then the total number of detected frames is reported. We apply the same method for the surveillance and security videos. Figures 8(a) and 8(b) compares the three video adaptation techniques for the 200 news and speech videos. These videos have collectively more than 2 million faces in their frames. The rate-accuracy curves are similar to those of the standard sequences, but the crossover point between spatial and temporal adaptations happens at much lower bitrates, and thus the spatial adaptation performs almost always better than the temporal. Moreover, the gap between SNR and spatial with upscaling becomes wider. This gap is somewhat exaggerated since the false positives are not considered. Recall that SNR tends to perform worse in the false positive index, as indicated by Figure 2. The overall patterns do not change much with the video compression scheme used.



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

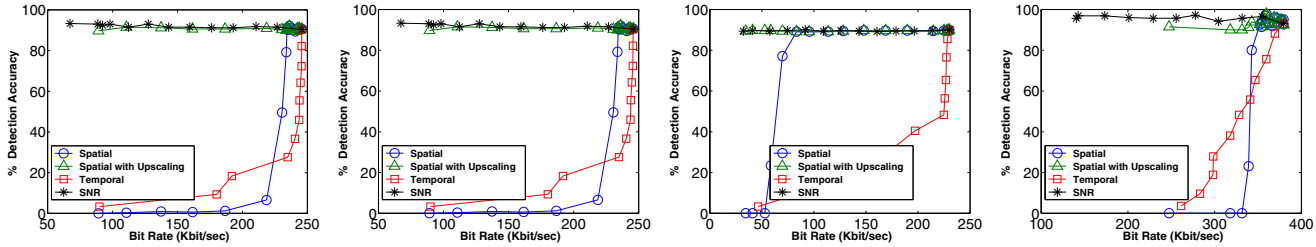
Fig. 3: Rate-Accuracy Curves for VTCLNW 4SIF Sequence [Upscaling done with Bicubic Interpolation]

Results for Surveillance and Security Videos: Surveillance and security videos are the closet to those that we would expect in AVS systems. Figures 8(c) and 8(d) compares the three video adaptation techniques for the 100 surveillance and security videos. These videos have collectively more than $60,000$ faces in their frames. Since the quality of these videos is generally lower than news and speech videos, all adaptations have somewhat worse rate-accuracy curves. In addition, the relative performance among different adaptation techniques remains unchanged, but the gap between temporal and spatial adaptations becomes narrower, and the gap between SNR adaptation and spatial with upscaling becomes significantly wider. As explained earlier, the latter gap is somewhat exaggerated.

Comparing Video Rate Adaptation Techniques in Power Consumption

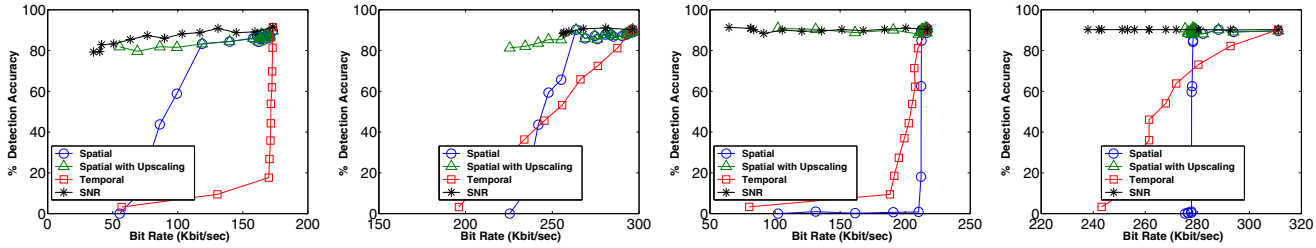
In AVS applications, power consumption at the video sources is usually considered as a primary concern because these sources may be battery-operated video cameras or sensors. The power consumption in the capturing phase generally depends linearly on the total number of pixels in the video [10], which is equal to the frame rate times the number of pixels in each frame. Thus, spatial and temporal rate adaptations are expected to require lower capturing power consumption than the SNR. The power consumption in the transmission phase depends on the achieved video bitrate. The power consumption in the encoding phase is the most significant.

Through actual experiments, we compared the three video rate adaptation techniques in terms of power consumption



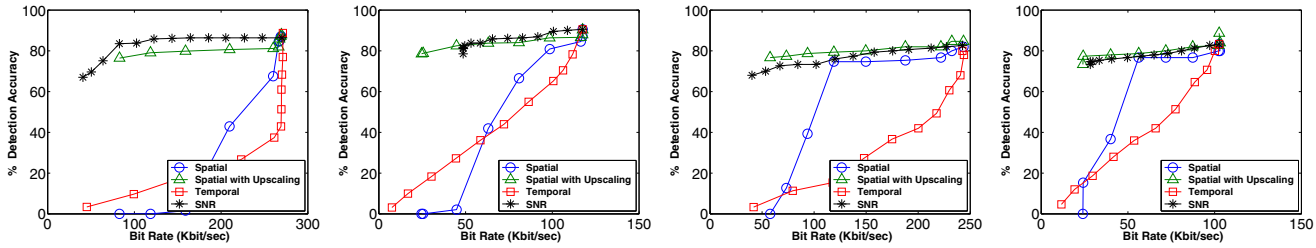
(a) Silent Sequence, MPEG-4 (b) Silent Sequence, H.264 (c) Akiyo Sequence, MPEG-4 (d) Akiyo Sequence, H.264

Fig. 4: Rate-Accuracy Curves for Silent and Akiyo CIF Sequences [Upscaling done with Bicubic Interpolation]



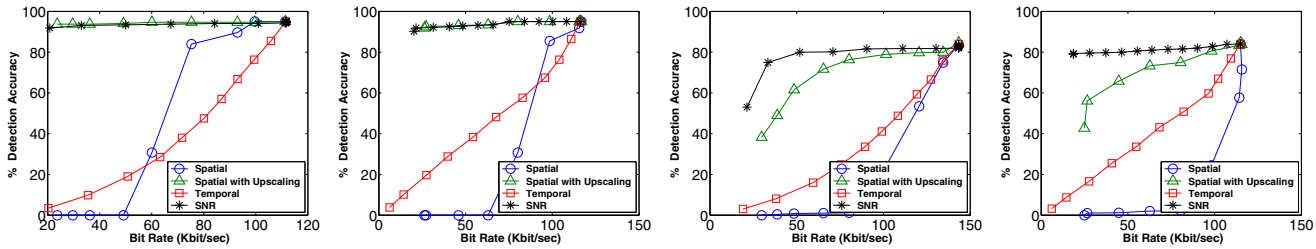
(a) Deadline Sequence, MPEG-4 (b) Deadline Sequence, H.264 (c) SignIrene Sequence, MPEG-4 (d) SignIrene Sequence, H.264

Fig. 5: Rate-Accuracy Curves for Deadline and SignIrene CIF Sequences [Upscaling done with Bicubic Interpolation]



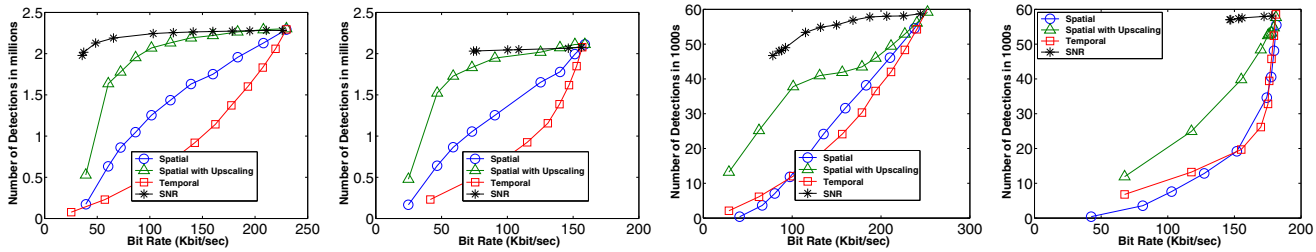
(a) Carphone Sequence, MPEG-4 (b) Carphone Sequence, H.264 (c) Suzie Sequence, MPEG-4 (d) Suzie Sequence, H.264

Fig. 6: Rate-Accuracy Curves for Carphone and Suzie QCIF Sequences [Upscaling done with Bicubic Interpolation]



(a) Akiyo Sequence, MPEG-4) (b) Akiyo Sequence, H.264 (c) Claire Sequence, MPEG-4 (d) Claire Sequence, H.264

Fig. 7: Rate Accuracy Curves for Akiyo and Claire QCIF Sequences [Upscaling done with Bicubic Interpolation]



(a) MPEG-4, 200 News Videos (b) H.264, 200 News Videos (c) MPEG-4, 100 Security Videos (d) H.264, 100 Security Videos

Fig. 8: Rate Accuracy Curves [Average Results of 200 News Videos "(a) and (b)", and 100 Security Videos "(c) and (d)", Upscaling done with Bicubic Interpolation]

in the encoding stage. Figure 9 shows that SNR rate adaptation results in the most power consumption, whereas spatial and temporal lead to significantly lower power consumption. Spatial and temporal adaptations are expected to lead to even lower overall power consumption as they also reduce the power consumption in the capturing phase. The power consumption results will vary with the implementation, but the general behavior will not change as long as the hardware uses Dynamic Voltage Scaling. Although these results are for H.264 compression, MPEG-4 exhibits a similar behavior.

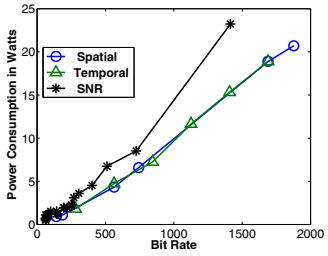


Fig. 9: Comparing Rate Adaptation Techniques in Power Consumption in Video Encoding [H.264 Encoding, FFmpeg Implementation, Default Resolution: 768×480, Default Frame Rate: 30 fps, Default QP: 22]

In video streaming systems in general and AVS systems in particular, the power consumption at the receiver is another concern. In AVS systems, the power is consumed by decoding the videos and running the computer vision algorithms and possibly the super-resolution algorithm. As the receiver may have to process a huge number of incoming streams (from video cameras and/or video sensors), power consumption at the receiver becomes a main concern. Thus, we should consider the additional power consumed by running the upscaling algorithm in the case of spatial adaptation.

5. CONCLUSIONS

We have analyzed the rate-accuracy characteristics of four video adaptation techniques (spatial, spatial with upscaling, temporal, and SNR), considering nine standard sequences and a total of 300 real surveillance, security, news, and speech videos. We have analyzed each video in both MPEG-4 and H.264 codecs, at 48 different quality levels. The results show that SNR adaptation generally achieves the best rate-accuracy characteristics, followed by spatial with upscaling, but the latter performs better in the false positive index. We have compared the performance of five upscaling algorithms. The results show that upscaling provides an outstanding improvement in the detection accuracy, but various upscaling algorithms perform close to one other. The Bicubic algorithm provides the best compromise between accuracy and complexity.

Power consumption is becoming a major concern, and thus it should be considered as well. We have analyzed the rate-energy characteristics of spatial, temporal, and SNR video rate adaptations by conducting actual experiments. The results show that SNR adaptation leads to significantly higher power consumption than spatial and temporal adaptations. When power consumption at the video sources is a primary concern (such as in AVS systems with battery-operated video cameras and/or sensors), spatial adaptation with later upscaling at the receiver is a good choice as it provides close performance to SNR in terms of detection accuracy but with much lower power consumption. A combination of spatial and SNR adaptations with later upscaling at the receiver might be the best choice for significantly reducing the bitrate and power consumption without a considerable negative impact on accuracy.

6. REFERENCES

[1] A. Eleftheriadis and D. Anastassiou, "Constrained and general dynamic rate shaping of compressed digital video," in *Proceedings of the 1995 International Conference on Image Processing*, 1995.

[2] Minjung Kim and Yucel Altunbasak, "Optimal dynamic rate shaping for compressed video streaming," in *Proceedings of the First International Conference on Networking-Part 2 (ICN)*, 2001, pp. 786–794.

[3] J. Kim, Y. Wang, and S. Chang, "Content-adaptive utility-based video adaptation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, July 2003, pp. 281–284.

[4] Cheng hsin Hsu and Mohamed Hefeeda, "Video quality for face detection, recognition, and tracking," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 7, no. 1, January 2011.

[5] Pavel Korshunov and Wei Tsang Ooi, "Critical video quality for distributed automated video surveillance," in *Proceedings of the 13th ACM International Conference on Multimedia*, Nov. 2005.

[6] Pavel Korshunov and Wei Tsang Ooi, "Video quality for face detection, recognition, and tracking," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 7, no. 3, August 2011.

[7] Yousef O. Sharrab and Nabil J. Sarhan, "WAV-Image Dataset," 2012, Dataset available at http://www.ece.eng.wayne.edu/ nabil/wavimage.html.

[8] Paul Viola and Michael Jones, "Robust real-time face detection," in *Proceedings of the ICCV 2001 Workshop on Statistical and Computation Theories of Vision (ICCV)*, July 2001, p. 747.

[9] Andrew King, "A survey of methods for face detection," Technical report, University of Toronto, March 3 2003.

[10] A. Dupret, J.O. Klein, and A. Nshare, "A programmable vision chip for CNN based algorithms," in *Proceedings of IEEE International Workshop Chip for Cellular Neural Networks and Their Applications Proceedings*, 2000.