MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Temporal Perceptual Coding Using a Visual Acuity Model

Adzic, V.; Cohen, R.A.; Vetro, A.

TR2014-006 February 2014

Abstract

This paper describes research and results in which a visual acuity (VA) model of the human visual system (HVS) is used to reduce the bitrate of coded video sequences, by eliminating the need to signal transform coefficients when their corresponding frequencies will not be detected by the HVS. The VA model is integrated into the state of the art HEVC HM codec. Compared to the unmodified codec, up to 45% bitrate savings are achieved while maintaining the same subjective quality of the video sequences. Encoding times are reduced as well.

IS&T / SPIE Symposium on Electronic Imaging

Temporal perceptual coding using a visual acuity model

Velibor Adzic*[a], Robert A. Cohen[b], Anthony Vetro[b]
[a]Florida Atlantic University, 777 Glades Road, EE408B, Boca Raton, FL, USA 33431
[b]Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA, USA 02139

ABSTRACT

This paper describes research and results in which a visual acuity (VA) model of the human visual system (HVS) is used to reduce the bitrate of coded video sequences, by eliminating the need to signal transform coefficients when their corresponding frequencies will not be detected by the HVS. The VA model is integrated into the state of the art HEVC HM codec. Compared to the unmodified codec, up to 45% bitrate savings are achieved while maintaining the same subjective quality of the video sequences. Encoding times are reduced as well.

Keywords: Perceptual Video Coding, Human Visual System, Visual Acuity, Subjective Quality, HEVC

1. INTRODUCTION

Psychophysical studies of the human visual system (HVS) revealed that through a very sophisticated and complex cascade of sub-systems it filters input information in such a way that only "important" components are kept. What this effectively means is that significant portions of the visual signal presented to our eyes never reach higher processing levels in our brain. It is not before the received signal passes through the HVS that the description of its quality can be given.

In general, for image and video coding systems, in order to obtain better quality we have to spend more bandwidth or use more bits to describe the original signal. However, having limited resources requires us to spend bits in a careful and optimized manner. Spreading bits uniformly across spatial and temporal dimensions of a compressed signal is often redundant because of the way the HVS processes visual stimuli. If we can filter out the same information that is not kept by the HVS we can achieve visually lossless compression while reducing the bitrate or size of the compressed data. For visually lossy compression, perceptual techniques also can be applied to obtain coding gains over non-perceptual coders.

While some of the aspects of the HVS have been considered in the design of the modern hybrid coders, many other characteristics are not exploited. Although newly released video standards such as HEVC[1] provide significant improvements over state-of-the-art codecs such as H.264/AVC[2], the reference encoders still make decisions based on non-perceptual metrics such as sum of squared error. Recent research has incorporated perceptual distortion models into these codecs, in which the quantization process is modified so that coarser quantization can be used in areas where the HVS is less sensitive to spatial distortion[3]. While reducing the overall bitrate, such methods do not eliminate the need for certain data to be signaled to the decoder. Furthermore, the existing perceptual coding models do not incorporate the concept of visual acuity as affected by motion. As described in the next section, research has shown that the sensitivity of the HVS to objects containing higher frequency components is affected by the velocity of the objects.

This paper presents a method for using a temporal visual acuity model in a video coder to eliminate the need to signal higher-frequency transform coefficients. The end effect is that blocks corresponding to areas with faster motion are coded using fewer coefficients, thus reducing the overall bitrate of the compressed data. Experimental results are shown for when this model is integrated into HEVC.

2. TEMPORAL VISUAL ACUITY MODEL

For the visual acuity model we are using results from the experiments by Kelly[4] and Eckert and Buchsbaum[5]. Kelly established the thresholds of contrast sensitivity at different velocities of moving sine-wave gratings. His results for the maximum spatial frequency resolved by the HVS for a fixed low contrast level are used to establish a velocity threshold.

*This work was performed while at Mitsubishi Electric Research Laboratories

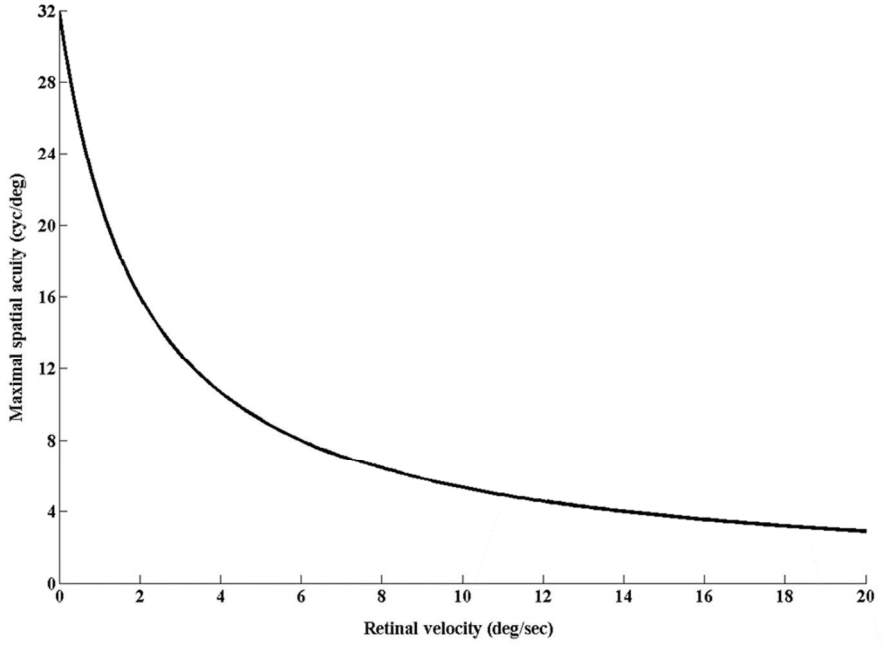

Figure 1. The maximal spatial acuity of a grating moving at a specified retinal velocity. The model assumes 32 cycles/deg to be the highest perceptible frequency of a stationary object. When objects move at high velocities this frequency is significantly reduced.

After separation of the model for horizontal and vertical components we obtain a maximum perceptible frequency

$$K_i = \frac{K_{\max} \cdot v_c}{v_{R_i} + v_c}, \quad (1)$$

where i denotes the x or y component, K_{max} is the highest perceptible frequency (32 cycles/deg in our paper) for a static stimulus, v_{Ri} is the x or y component of the retinal velocity of a stimulus and v_c is the corner velocity, i.e. where the spatiotemporal sensitivity of the HVS is greatest, fit to Kelly's data. In this model $v_c = 2$ deg/s.

A reasonable estimate for the number of pixels per degree of viewing angle under typical high-definition viewing conditions is 64 pixels/deg. To avoid the need to signal additional side information indicating the velocity of objects, we use motion vectors as an approximation, as the decoder already has motion vectors available prior to decoding the coefficients.

The velocity on the image plane of a block can then be computed given its motion vector length and the frame-rate of the video. We use a model from Daly[6], which computes the eye velocity as a function of target velocity on the image plane and then subtracts the eye velocity from image plane velocity to obtain retinal velocity v_R.

A graph of the maximal spatial acuity expressed as a function of object's retinal velocity is shown in Figure 1. We can see that as an object starts moving with higher retinal velocity, the highest perceptible frequency diminishes fairly quickly. While high retinal velocities (above 4 deg/sec) are not common for moving objects in real life, some video sequences contain objects or regions that have high velocities as a result of composition and camera movement. Our model allows a significant reduction of transmitted spatial frequency information in those cases.

For each block of transform coefficients that have an associated motion vector, the value K_i is used to compute *horizontal* and *vertical* position thresholds, for which all transform coefficients located in rows or columns above those thresholds are truncated, i.e. they are not signaled in the bit-stream.

The thresholding is equivalent to setting the removed coefficients to zero. An example of this coefficient thresholding is depicted in Figure 2. Here we have an 8x8 transform block, but the same principle is extensible to any block size. Using the relationship between maximum perceptible frequency K_i and velocity v_{Ri} expressed in Equation (1) and illustrated in Figure 2(c), we can determine the number of columns or rows that can be removed from the horizontal and/or vertical

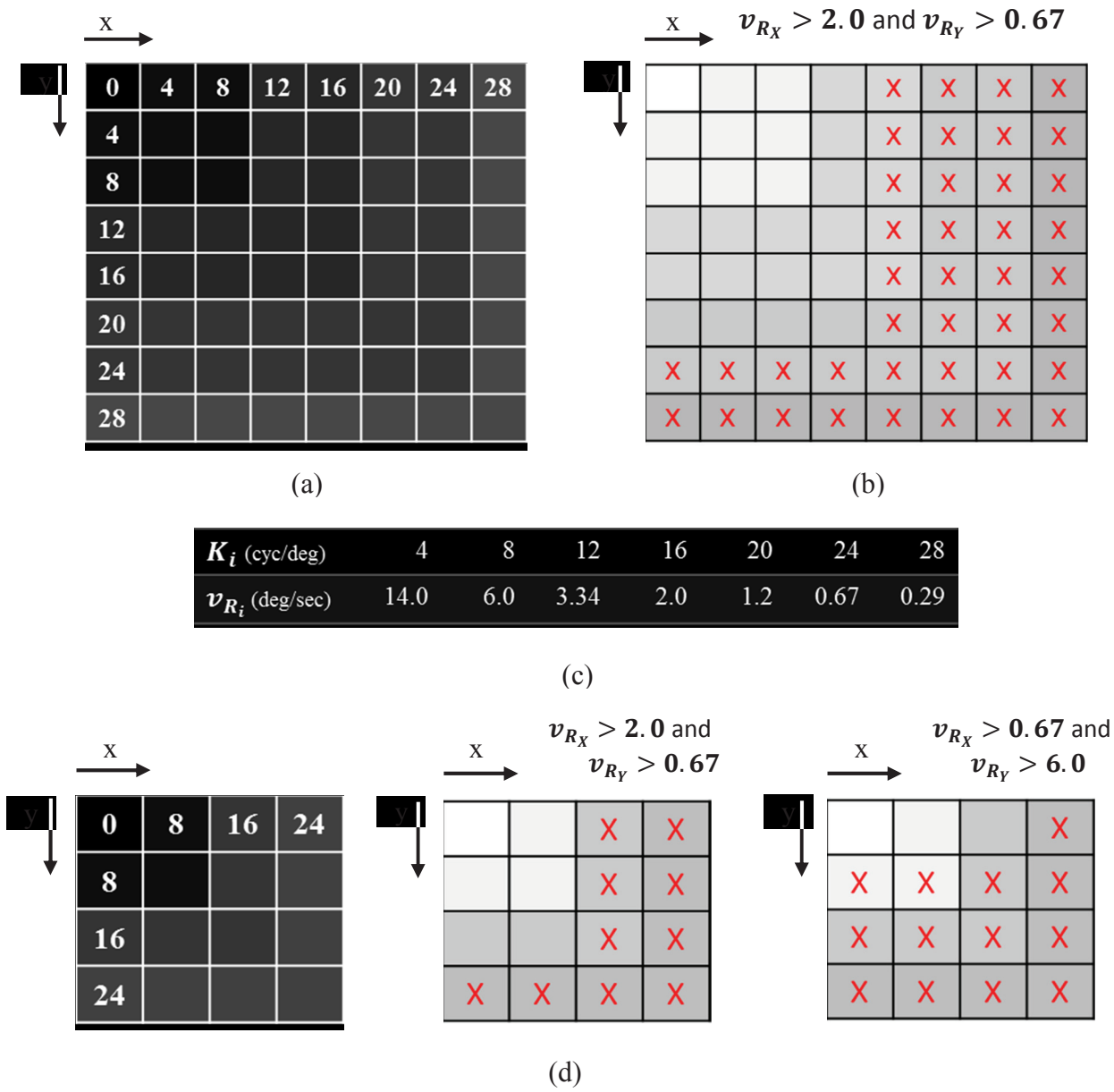

Figure 2. Examples of coefficient truncation. (a) 8x8 block of transform coefficients with designated spatial frequency bands for horizontal and vertical components. (b) Example of coefficient truncation based on horizontal and vertical retinal velocity of the moving object that is represented by this 8x8 transform block. Coefficients that are covered with "X" are truncated. (c) Table showing velocities associated with a range of maximum perceptible frequencies. (d) Thresholding examples for a 4x4 transform block.

components in the frequency domain. In this example, the horizontal component of the velocity is 2.0 deg/sec, which corresponds to 16 cycles/deg. Given that 16 cycles/deg is half of the K_{max} value of 32 cycles/deg, the model flags the four columns corresponding to high horizontal frequencies as being imperceptible. The vertical threshold corresponds to 24 cycles/deg, which means that the two highest-frequency rows are marked as imperceptible. Examples for 4x4 blocks are shown in Figure 2(d).

Other aspects of implementing this method in the HEVC encoder and decoder are as follows:

- The distortion used in cost functions is modified so as not to be penalized by the removal of these coefficients, according to the perceptual model.

- HEVC uses a data hiding method to embed the sign bit information into the signaled transform coefficients. The coefficients that are truncated via our perceptual coding method are ignored, i.e. skipped over, during the data hiding process.

- All truncated coefficients are skipped by the entropy encoder and are generated by the entropy decoder as zero-valued coefficients. No additional side information needs to be signaled because the perceptual model uses motion vector information to derive the location of truncated coefficients.

3. EXPERIMENTAL RESULTS

Subjective tests were conducted in order to verify the model. Tests were done according to the BT.500 recommendation[7]. A total of 10 subjects (8 male, 2 female) completed the experiments over the course of one week. Subjects were volunteers, between 20 and 35 years old, with normal or corrected to normal vision. All test sequences were presented on a 24 inch LCD monitor that supports Full HD (1080p) playback.

Test sequences were selected as a subset of those commonly used for HEVC development. We chose 6 sequences from different classes of the dataset, as shown in Table 1.

Table 1. Test sequences used for subjective experiments.

Sequence	Width (px)	Height (px)	Frame rate (fps)	Duration (s)
BQTerrace	1920	1080	60	10
FourPeople	1280	720	60	10
BasketballDrill	832	480	50	10
PartyScene	832	480	50	10
BasketballPass	416	240	50	10
BlowingBubbles	416	240	50	10

The HM 11.0[8] software was modified to incorporate the temporal visual acuity model. The unmodified codec, denoted as “HM”, was run for a given quantization parameter (QP) value, and then the modified codec “VA” was run using the same QP. Values for QP were selected to fit different scenarios: A QP value of 10 is close to visually lossless, a QP value of 30 may be suitable for Internet streaming, and value of 20 is selected as a middle ground test case.

The methodology used for comparison is Stimulus Comparison Adjectival Categorical Judgement[7] (ACJ). Sequences are presented as randomized pairs after which the subject is asked to assign a score to the sequence that is second in order. The score is a measure of the quality of a second sequence as compared to a first sequence. It ranges from -3 which is interpreted as “much worse” to +3 which is interpreted as “much better”. The score 0 means that the video sequences have identical quality. Each subject viewed all pairs of sequences in a random order. This ensured that all sequence pairs were assessed by all subjects, hence removing possibility of a per-subject or per-sequence bias.

The results were obtained through ANOVA analysis and calculated for a 95% confidence interval. As shown in Figure 3, these results show that the subjects did not notice any differences between the sequences.

In the tests pairs of VA and HM sequences are presented in a random order, but the results in Figure 3 always show the quality of the VA sequence relative to the HM sequence. As can be seen, the average scores are positive across all subjects and very close to 0 for all sequences. The interpretation of results validates the claim that subjects are not able to perceive any difference in quality between the original (HM) and modified (VA) sequence.

The Bitrate savings that are achieved by using our model can be seen in Figure 4. In order to better understand difference in achieved savings we are also showing the spatio-temporal (ST) information of all test sequences in Figure 5. The ST-information was calculated according to P.910 recommendation[9].

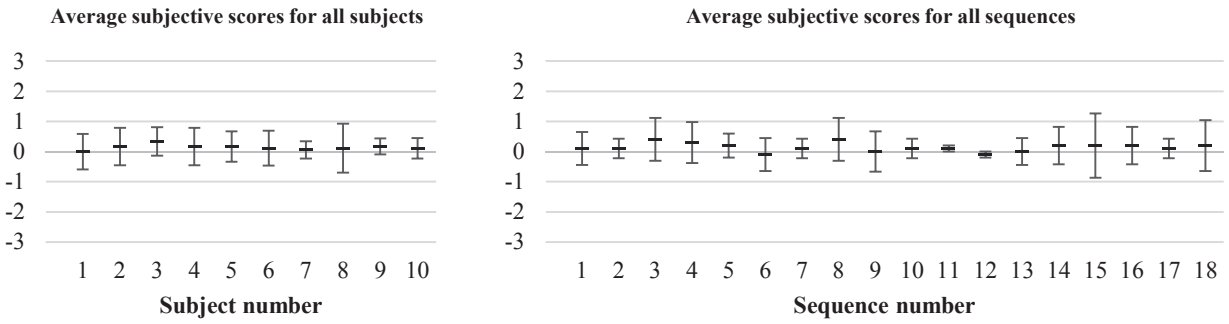

Figure 3. ACJ scores obtained from experiments using ANOVA: Left graph shows average ACJ scores per subject for all sequences; right graph shows average ACJ scores per sequence for all subjects. Scores are presented with error bars. Sequence numbers are shown in Table 2

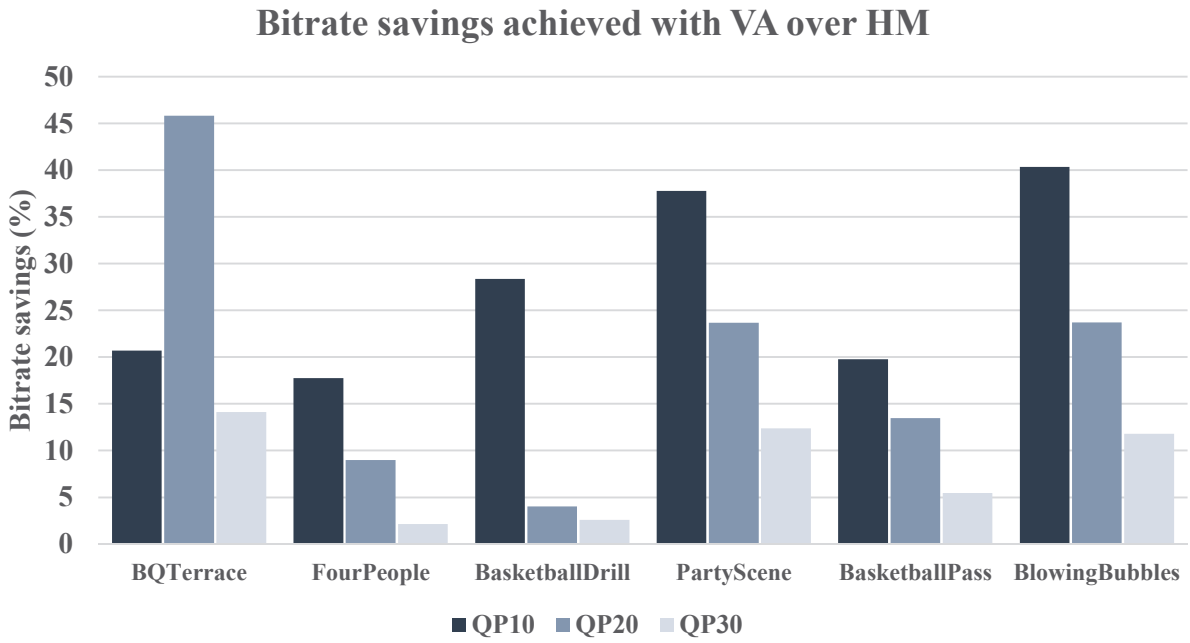

Figure 4. Bitrate savings calculated as a percentage decrease in achieved bitrate for VA sequences compared to HM.

For the vertical Temporal Index axis of Figure 5, low values correspond to sequences having very limited motion. High values indicate that a sequence contains scenes with a lot of motion. For the horizontal Spatial Index axis, low values correspond to scenes having minimal spatial detail, and high values are found in scenes having significant spatial detail. Obviously, the savings achieved with our method are dependent on the motion activity and the amount of high frequency components in the spatial domain. The sequence *FourPeople* has a very low temporal and spatial information score and is also the sequence for which we achieved the least amount of overall savings. It is worth noting that beyond some point the ST-information is not a good predictor of savings when using the VA model because the way motion activity is calculated for temporal information is different from the way the encoder implements motion estimation. Overall, the ST-information graph also shows that our dataset covers broad spectrum of content.

Detailed data for bitrate and time savings are presented in Table 2. As can be seen, when implementing the VA model into HM we are able to achieve not only bitrate reduction but also a reduction in encoder complexity, as reflected by encoder run-times. Reductions in total encoding times of up to 40% are achieved.

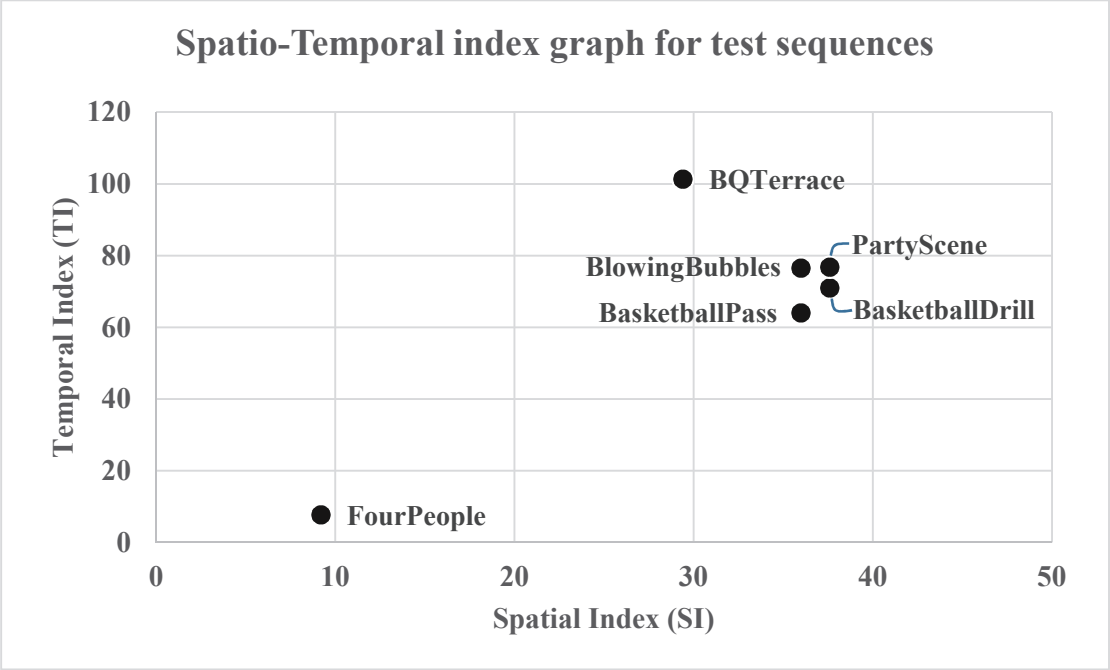

Figure 5. Spatio-temporal information plot for all test sequences, calculated according to P.910 recommendation.

Improvements in both compression efficiency and in encoding times are also related to QP. In most cases the VA method exhibits more improvement as QP is decreased because smaller QP values produce more nonzero high-frequency coefficients, which are subsequently removed by the VA method, assuming there is sufficient motion in the scene. Although we chose HM (HEVC) as a reference encoder, the model is sufficiently simple and robust to allow implementation in any encoder that uses frequency domain transforms.

Table 2. Bitrate and time savings achieved by using VA model, compared to HM.

No.	Sequence name	QP	HM Rate (kbps)	VA Rate (kbps)	Rate reduction (%)	Encode time reduction (%)
1	*BQTerrace*	10	346884.318	275148.653	20.68	34.30
2		20	98135.518	53176.883	45.81	35.05
3		30	4423.675	3799.109	14.12	2.31
4	*FourPeople*	10	65685.362	54037.766	17.73	27.63
5		20	4949.805	4504.677	8.99	4.15
6		30	699.12	684.125	2.14	-4.43
7	*BasketballDrill*	10	34747.602	24898.366	28.35	35.31
8		20	5793.206	5559.678	4.03	21.52
9		30	1271.187	1238.166	2.60	6.76
10	*PartyScene*	10	58101.914	36164.006	37.76	40.92
11		20	17325.12	13224.464	23.67	34.29
12		30	3363.851	2947.418	12.38	15.81

13	*BasketballPass*	10	6114.064	4904.914	19.78	29.65
14		20	2021.907	1749.283	13.48	23.54
15		30	497.494	470.381	5.45	10.60
16	*BlowingBubbles*	10	13365.958	7972.954	40.35	40.14
17		20	3126.962	2385.389	23.72	27.66
18		30	554.278	488.906	11.79	8.41

Examples of decoded HM and VA frames for *PartyScene* coded with QP=30 are shown in Figure 6. The left side of the frame region contains very fast motion around the person's legs. The right side contains stationary objects, so the only motion is due to the slow panning of the camera. In the VA decoded frame, it can be seen that there is more distortion around the legs as compared to the HM decoded frame. These differences are easily visible in a still frame, but when viewed at the full frame-rate, the HVS does not perceive the distortion due to the high speed of the motion. The HM and VA decoded frames are similar in quality on the right side of the frame, as the VA model truncates fewer or no coefficients due to the very slow panning motion.

4. CONCLUSIONS

This paper summarizes work showing how a temporal visual acuity model can be used to improve the coding performance of HEVC by eliminating the need to signal coefficients based upon the frequency content and velocity of blocks. Using motion vectors as an estimation of the horizontal and vertical motion of a block, the visual acuity model is used to compute frequency thresholds. Coefficients corresponding to spatial frequencies above the thresholds are not signaled in the bit-stream. Reductions in bitrate of up to 45% were obtained, and savings in encoder run times of up to 40% were achieved. Future research will include combining it with spatial perceptual models, which have been shown to produce additional gains in compression efficiency.

REFERENCES

[1] Sullivan, G. J., Ohm, J., Han, W.-J. and Wiegand, T., "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst.Video Technol.*, **22**, pp. 1649–1668 (2012).

[2] Wiegand, T., Sullivan, G. J., Bjontegaard, G. and Luthra, A., "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, **13**, pp. 560–576 (2003).

[3] Naccari, M. and Pereira, F., "Integrating a spatial just noticeable distortion model in the under development HEVC codec," in *Proc. International Conf. on Acoustic Speech and Signal Proc. (ICASSP)*, Prague, Czech Republic (May 2011).

[4] Kelly, D. H., "Motion and vision. II. Stabilized spatio-temporal threshold surface," *Journal of the Optical Society of America* **69**, no. 10, pp. 1340–1349 (1979).

[5] Eckert, M. P. and Buchsbaum, G., "The significance of eye movements and image acceleration for coding television image sequences," in [*Digital images and human vision*], MIT Press, pp. 89–98 (1993).

[6] Daly, S. "Engineering observations from spatiovelocity and spatiotemporal visual models," *SPIE HVEI III* **3299**, pp. 180–191 (1998).

[7] Recommendation ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva (2012).

[8] HM-11.0-dev software. https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/branches/

[9] Recommendation ITU-T P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," International Telecommunication Union, Geneva (1999).

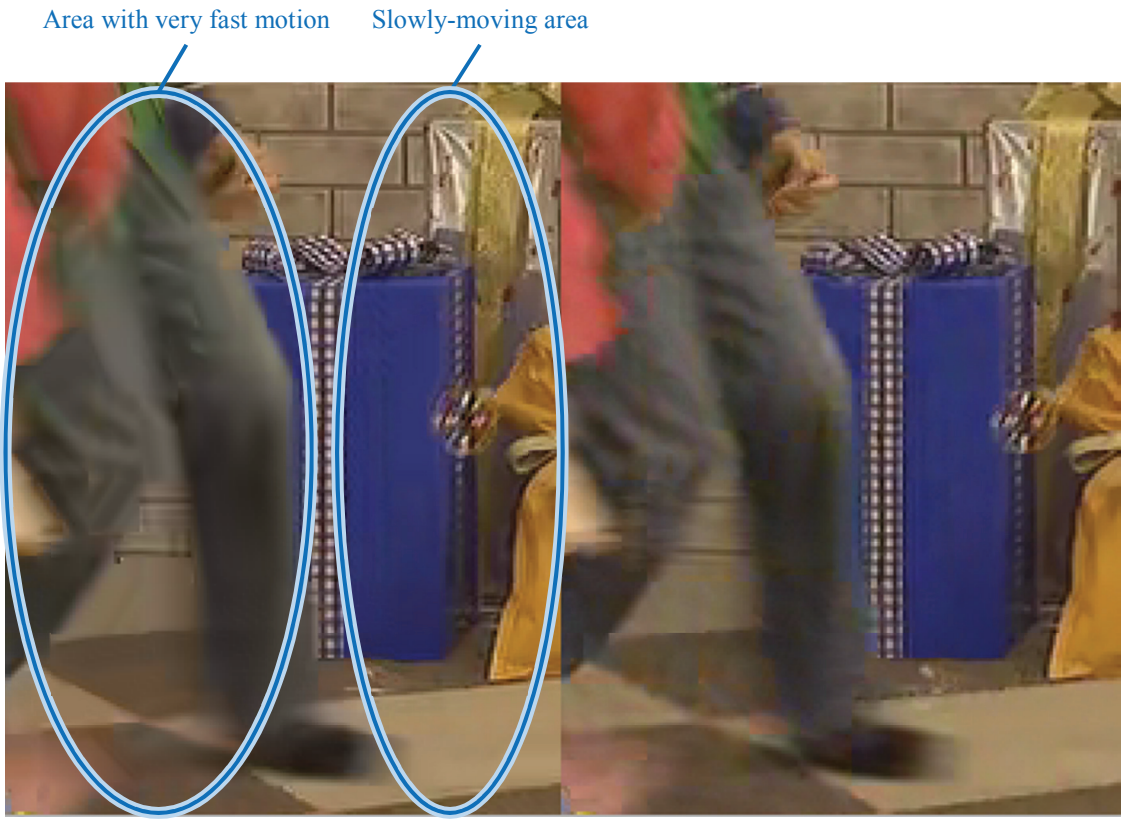

Figure 6.Magnified region of a decoded *PartyScene* frame containing significant differences in motion. Both the HM and VA frames were coded using QP=30.