Low complexity sub-block perceptual distortion assessment for mode decision and rate-control

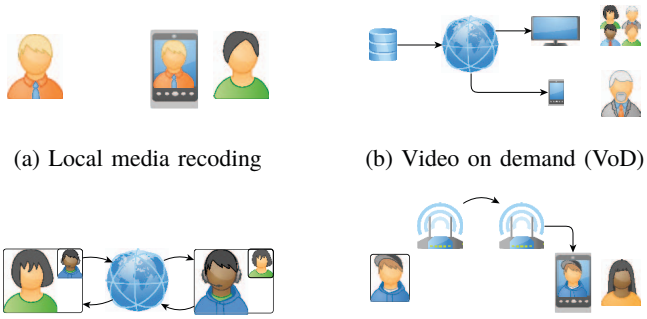
Y. G. Joshi, J. Loo, P. Shah, S. Rahman, and A. Tasiran
School of Science and Technology, Middlesex University, The Burroughs, Hendon, London, NW4 4BT, UK
Email: {y.joshi, j.loo, p.shah, s.rahman, a.tasiran}@mdx.ac.uk

***Abstract*---Video is being used in a variety of portable and low powered devices, via online/on-demand services, for personal communications or over heterogeneous/wireless sensor networks. Likewise, perceptual evaluation is being sought, to ensure video sequences maintain perceptual integrity. This raises a challenge, to bring high complexity perceptual algorithms into a low complexity environment. Existing perceptual solutions minimise the overall complexity by making the Lagrange multiplier (λ), the quantisation stage perceptually aware. These solutions are restricted to the block level using the original pixels, a model or previous encoded block, thus avoiding assessing individual sub-block candidates. Current perceptual algorithms like structured similarity (SSIM) uses statistical based calculations, and in order to be compatible with existing scores a further high complexity function is required for scaling. This paper presents a perceptual distortion and activity assessment that can operate at the sub-block for each candidate during the later stages of encoding, in mode-decision and in rate-control, without the need for statistical calculations nor the high complexity associated with scaling a perceptual algorithm. The paper will show how several perceptual techniques of SSIM luma function, just noticeable detection (JND) and a new proposed edge detection can be used to form a low complexity solution. Consequently, the proposed low perceptual assessment has additional timing increase of < +4 % for medium and low activity video sequences.**

I. Introduction

Video is increasingly generated by and delivered to low powered devices, which increases the mobility and accessibility of video. The traditional demands on the infrastructure of bandwidth remain, with the added factors of power consumption. Likewise, innovation is being sought for incorporating perceptual video coding (PVC) so that video encoding may retain perceptually significant features. As PVC is highly complex there is a challenge to make PVC within a low complexity envelope. The applications for such a solution are broad and can be placed into four major applications as shown in figure 1 of local storage, on-line/on-demand, personal video communications and wireless sensor networks. In fact, these applications can be drawn as storage critical or responsive critical. Figures 1a and 1b is where video is being stored or accessed for later period in time, which requires efficient use of available storage space. Figures 1c and 1d are where video depends on responsiveness and/or expects the network to incur dropped frames, meaning the video must not depend on any one frame.

In video coding, a frame is made from a mosaic of blocks, of different sizes, the choices of which reflect the balance between representing the content and regulating bandwidth.



(a) Local media recoding (b) Video on demand (VoD)

(c) Personal video comm. (d) Wireless sensor networks

Fig. 1: Applications for low complexity perceptual video coding

These square block sizes under high efficiency video coding (HEVC) can vary from 16x16 to is 64x64 and called the largest coding unit (LCU). The sub-block is any square size less than the LCU, represented as 2^n where n is a minimum of 2. Under rate-control, a fixed size of 8x8 is analysed for activity, while in mode-decision square sizes down to 4x4 are assessed for distortion. This is illustrated in figure 2 which leads to the choice of distortion metrics in the encoder. In the case of rate-control, a fixed 8x8 size variant of the Hadamard transform is applied, while for mode decision the block sizes can vary, leading to the use of a pixel based distortion assessment of sum of square errors (SSE).

Video encoding is about finding those combinations of sub-blocks or a single block for a given bit-rate constraint which can produce the minimum energy of distortion. For rate-control this means being able to adjust the quantisation to meet the bit-rate constraint. This is known as the Lagrange

(a) Mode decision: varying size, non-uniform distribution (b) Rate-control: fixed 8x8 size, uniform distribution

Fig. 2: 64x64: Sub-block size and distribution for mode decision (varying, non-uniform) and rate-control (fixed, uniform)

optimisation, where two opposing resources in this case rate and distortion is represented as a curve, here called an R-D curve, with the aim of finding the closest point to the origin on convex hull [1]. This is shown in equation (1)

$$J = D + \lambda \cdot R \tag{1}$$

where, J is energy of the (sub-)block, D is distortion based on the type of distortion metric used, λ is the optimal level of quantisation required to meet the constraint rate defined by R. Making video coding to be perceptually aware involves incorporating models based on the Human Visual System (HVS), so that features which are perceptually important are retained. This means identifying perceptually significant regions via a perceptual distortion assessment within mode decision so that smaller blocks are used where perceptually significant and larger blocks where perceptually homogeneous. Extending perceptual coding to activity assessment within rate-control will promote the bits budget distribution upon the partitioning or on the quantisation applied to the given block. Finally, as the distribution of bits and partitioning may change the overall frame measurements may remain the same, measuring these influences via the use of participants can be time-consuming. As such frame level measurements may be insufficient for where perceptual video coding affects the sub-block level, and more localised means are required to evaluate where and by what amount changes have occurred without the use of participants.

II. Literature Review

The human visual system (HVS) is regarded as a complex system, described as a series of stages. The initial stage has received much attention for several decades and models have been produced. However, only in recent years has there been interest on applying PVC on assessing the significance on the distortion than the content alone. Structured similarity (SSIM) is widely used as a perceptual measure of distortion, often cited as an alternative to peak-signal to noise-ratio (PSNR), and incorporated into block-based encoder as a form of PVC. The popular choice of SSIM as a perceptual assessment is because it is regarded as the least complex among its peers [2]. Unfortunately, when SSIM is compared to existing non-perceptual distortion metrics it is highly complex. There are solutions that incorporate SSIM and address the issue of complexity, they adapt when perceptual assessment occurs, from the frequent calls of mode-decision to the reduced calls of rate-control, which leads to a reduction in the overall complexity [3]. This means that at the rate-control stage a perceptual response model is produced by applying varying degrees of quantisations and calculating a curve of best fit. As such, a perceptual R-D curve guides the encoder, re-scaling λ as a perceptual (λ_p) based on non-perceptual distortion (D) during the mode decision stage. Furthermore, since SSIM scores are not compatible with existing scores, there is the need to scale within the same distortion range, often cited as the triangle inequality ($\triangleq$) problem with SSIM [4]. This additional stage of scaling can be even more complex than SSIM itself, which further compounds the issue of PVC being unattractive to low powered devices and leading to two major issues:

1) As a means to save complexity cost, λ as λ_p relies upon non-perceptual distortion assessment based upon pixel differences, avoiding the perceptual significance of the original pixel values.
2) The technique of λ_p means that perceptual distortion during rate-control activity is assessed ahead of mode decision at the frame or block level, thus avoiding perceptually assessing each mode decision candidate individually.

These limitations of perceptual quantisation mean that a perceptual distortion (D_p) solution which replaces D in equation (1) can potentially address these issues. Previous work into the behaviour of perceptual verses non-perceptual at sub-block level illustrates SSIM and existing non-perceptual distortion metrics, evaluate distortion differently [5]. Upon further investigation, covariance, a component of SSIM, was shown to correlate well with the perceptual model of Just Noticeable Distortion (JND), an insight that had not previously been presented [6]. This lead to the non-linear scaling of SSIM without the need of highly complex mathematical operations of logarithms or exponentials as provided by other SSIM scaled solutions. Despite this, the complexity of SSIM makes the solution unattractive at the sub-block level.

The approach by existing perceptual quantisation solutions have considered them from the block size, the LCU, ignoring the different sub-block structure as shown in figure 2. In terms of mode decision, the perceptual distortion assessment must support the different sub-block sizes as well as the various severity of distortion, leading to the use of SSE, pixel based distortion assessment [7]. With regards to rate-control, HEVC has replaced the mean absolute difference (MAD) distortion assessment with a variation of the Hadamard transform, a fixed 8x8 size without the DC value [8]. Consequently, any perceptual solutions will need to operate within the distortion metric space and the mode of operation of these existing non-perceptual measures to minimise complexity overhead. This means in mode decision a pixel-based perceptual solution and in rate-control a solution that can complement the existing Hadamard transform.

Typically, perceptual algorithms from imaging domain have been evaluated by participants, and in recently years image databases have used to evaluate different perceptual image quality assessments (IQAs) [2], [9]. This gives credence to incorporate perceptual IQA such as SSIM into tools to evaluate video encodings. However, the use of perceptual IQAs in the video domain on the decoded frame is limited. Video use of spatial and temporal techniques means that the search for minimum energy (J) as described in equation (1) is subject to managing bandwidth. This means that the encoder must dynamically adapt to the changing bit-target and result in different signalling choices. As these signalling choices are

governed by the bit allocation of rate-control and the search of J in mode decision this can affect the partitioning and level of quantisation. Visualising these signalling changes on the decoded frame can inform as to where the encoder is allocating bits and how broad or narrow the sub-block partitioning is to represent the content. Currently, commercial tools offer this feature, but non-commercial are limited to decoded video. Such a tool can be extended to support the development of new perceptual algorithms as a means to simulate or even verify its behaviour.

III. Research Design

PVC and low complexity can seem at odds with each other, as their is a risk to the robustness of any perceptual solution. Since perceptual IQA involves error normalisation, where differences are considered in their perceptual significance, finding methods to minimise complexity during normalisation is crucial. However, the use of perceptual IQA is not suitable for all conditions, it should be reserved for those sub-blocks where it is beneficial. This means that D_p score is a combination of traditional non-perceptual IQA and perceptual IQA, where perceptual IQA is added to those candidates which have a perceptually poor distortion. As such, there is a requirement to identify those sub-blocks which have a perceptually poor distortion without the complexity of perceptual normalising the distortion. It is proposed that to reduce the overall perceptual complexity a sample of the given sub-block candidate is assessed, from which further perceptual IQA is considered, as illustrated in figure 3.

Another issue raised is that the incompatibility of perceptual IQA and non-perceptual IQAs which leads to further complexity. This is partly due to the use of windowing by perceptual IQAs like SSIM to average changes within a given 8x8 window. Addressing this issue means moving away from fixed size windowing perceptual IQA, and into pixel based perceptual assessment. The benefit is that processor friendly techniques can be designed into perceptual normalisation to minimise complexity. Unfortunately, with pixel based perceptual IQA there is a risk of losing perceptual integrity and so a secondary low cost perceptual stage is required to increase perceptual robustness. Any secondary perceptual stage should consider adjacent pixels, the rate of change in perceptual normalisation, in the form of an edge-detection. Current edge-detections are complex and cannot fit within the smallest sub-blocks, hence a new edge-detection is required.

IV. Methodology

As stated above and shown in figure 3 normalising the entire sub-block can be expensive, instead sub-blocks should be sampled from which checks are made to identify if full perceptual assessment should occur. This can ensure additional processing is applied where necessary and limiting the overall complexity. The first task is to perceptually account the distortion of a sample sub-block so it is perceptually normalised than by pixel difference. Choosing those pixel locations that make up the sample sub-block is dependent upon the tests

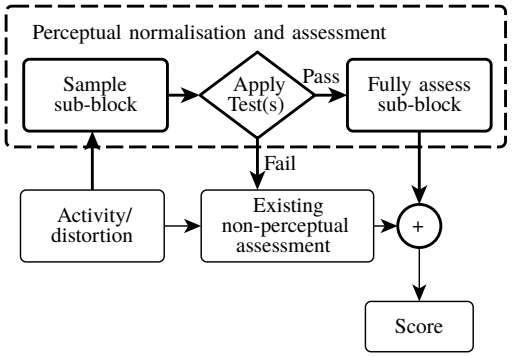


Fig. 3: Porposed perceptual normalisation and assessment workflow

to be applied. Here, a proposed test is where each side of the perceptually normalised version of the sub-block is taken (minus the corners) and subtracted from each other to find perceptually asymmetrical side labelled as PAS in equation (2).

$$PAS = (||T - B| - |L - R||) > Thresh \tag{2}$$

where T, B, L, R and $Thresh$ are top, bottom, left, right and threshold respectively. For mode-decision and rate-control the threshold values are 8 and 48 respectively, designed to reduce observations being processed by $\approx 50\%$, which are based upon analysis of 8x8 sub-blocks observations. Other specific tests to mode decision and rate-control are applied and are discussed below.

A. Mode Decision

Under mode decision the choices are between a combination of block sizes, a single block size or existing encoded block, whichever offers the least combined bit usage and distortion. At this stage of the encoder, each variation of the different block sizes represents the best prediction candidate, and so the distortion costs between these variations of block sizes may be minor. Whereas a traditional distortion metric seeks the minimum uniform cost, based on pixel difference, while a perceptual distortion assessment will consider the cost relative to the original pixel value. This means that the existing encoder workflow will need to be modified in order to pass the original pixel values at the last possible stage.

In the literature review, perceptual is calculated separately and not when existing distortion metrics are calculated. To reduce the overhead in perceptual processing, it should be integrated with non-perceptual, so the perceptual cost should be accumulated per pixel (like in non-perceptual) rather than averaged out across a block (like in SSIM). This means reducing the dependencies upon statistical calculations of mean, variance and covariance and using an in-line method to achieve similar forms of perceptual normalisation. While covariance is known to be a key perceptual component of SSIM, covariance in itself is still processor intensive. Instead, sum of square errors (SSE) should be adapted to support pixel based perceptually aware methods, in particular based upon the

calculation of covariance. Covariance involves both original (x) and reconstructed (y) pixels ($_i$), see equation (3).

$$\sigma_{x,y} = \frac{\sum (x_i \times y_i) - \Sigma x \times \mu_y}{n} \tag{3}$$

where $\sigma_{x,y}$ is covariance, μ is mean and n is the block size.

The sum product of original and reconstructed pixels can be rewritten as equation (4), where SSE is $(x-y)^2$.

$$\sum (x_i \times y_i) = \frac{\Sigma x_i^2 + \Sigma y_i^2 - SSE}{2} \tag{4}$$

It should be noted that in equation (4) the squared operation of original and reconstructed pixels can be represented by a look up table (LUT) to further reduce the complexity overhead. This allows covariance to be rewritten as equation (5).

$$\sigma_{x,y} = \frac{\frac{\Sigma x_i^2 + \Sigma y_i^2 - SSE}{2} - \Sigma x \times \mu_y}{n} \tag{5}$$

Unfortunately, equation (5) still involves one multiply and one divide for calculation of covariance, the divide must exist and cannot be substituted with a right shift. To reduce the complexity further while maintaining the perceptual properties, a new less intensive operation based on equation (4) is required.

1) Sum of square differences: During mode decision the variation between candidates can be very small, this means scores will need to be perceptually assessed at the pixel level. Combining a non-uniform cost with a uniform cost can risk destabilisation of the overall score, hence the perceptual cost should only represent a small proportion (within $\approx 10\%$). The proposed perceptual cost will be scaled down by a given factor depending upon its block size. This paper proposes SSE with Sum of Absolute Squared Differences (SASD), as shown in equation (6),

$$SASD = (\left|\left|x_i^2 - y_i^2\right| - SSE\right|)/2^8 \tag{6}$$

To ensure that SASD is used only where perceptual distortion activity is high, a threshold of $2^{(2n+3)}$ is applied, where n is $log_2(blockwidth)$, or as shown in table I. Also, since SASD is right-shifted by eight, the entire range of potential pixel costs can be stored in memory within a LUT with values between 0 and 255.

2) Edge detection: As described in section III a pixel based perceptual normalisation risks perceptual robustness and that another form of perceptual assessment is required. Here a new 2x2 edge detection is proposed, so that SASD is applicable where regions are textured.

Perceptual and non-perceptual are known to have similar correlation at either end of the scales, but differ in their response in between [10]. As such normalising distances in perceptual terms does not factor in the perceptual integrity of the block. To achieve this an edge detection should be applied on the perceptual normalised block. As most edge detections are large and cumbersome, a new proposed edge detection is presented in the form of a 2x2 sized edge detection as shown in equation (7),

$$Edge = (2 \cdot Centre) > (Top + Left) \tag{7}$$

where $Edge$ has a value of zero or one, and $Centre$, Top and $Left$ are pixel values. To keep the overall cost down four 2x2 edge detections will operate in a set pattern within every 4x4 block. As this edge detection is shaped as an `L', it can be orientated to test different pixels for perceptually significant boundary changes. The choice of this pattern of non-overlapping edge-detect provides coverage with minimum test points and can be referred to as 1/4 sub-block edge detection within mode decision. Finally, if sufficient edges are detected, then the block may add SASD cost to SSE as defined in table I.

B. Rate-control

During rate-control, block activity is used as a measure of how detailed the content is, and thus influence the number of bits to be allocated. This principle should be extended in perceptual terms, allowing bit allocation to be adjusted according to the perceptual activity. As rate-control activity employs a variant of Hadamard, any perceptual score should be applied in similar techniques to ensure consistent behaviour. This means that perceptual normalisation should be conducted under a similar pattern to Hadamard under rate-control, except that differences are perceptual and not pixel based. Since Hadamard will use pixel pairs of distances of 1, 2 and 4, the perceptual differences could be very large at times, making SASD unsuitable for perceptual activity under rate-control. A more advanced perceptual model is proposed which combines two perceptual models, SSIM luma function and just noticable difference (JND) background luminance masking. The SSIM luma function is described in equation (8), it is part of the trio of functions that eventually produce SSIM.

$$SSIM_l(x,y) = \frac{2\mu_x \times \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{8}$$

where C_1 is a constant based upon the maximum pixel range, [9]. While the JND background luminance masking is part of the JND perceptual measure and shown here in equation (9) [11], [12].

$$JND_l(x,y) = \begin{cases} 17 \times (1 - (\frac{bg(x,y)}{128})^{\frac{1}{2}}) + 3 & bg(x,y) \leq 127 \\ \frac{3}{128} \times (bg(x,y) - 127) + 3 & bg(x,y) > 127 \end{cases} \tag{9}$$

Threshold	4x4	8x8	16x16	32x32	64x64
Sum of squared differences	2^7	2^9	2^{11}	2^{13}	2^{15}
1/4 block edge detection	2^1	2^3	2^5	2^7	2^9

TABLE I: Mode decision - block size specific threshold values for both SASD and 1/4 block edge detection

Types of operations	Existing RC Had	Sample dom. side (stage 1)	Sub total	Cond. corner (stage 2)	Sub total	Diff. vs. RC Had	Proposed inside (stage 3)	Total	Diff. vs. RC Had	Altern. SSIM
Multiply, Divide	0	0	0	0	0	0	0	***0***	0	***208***
Addition, Subtract.	577	78	655	18	673	96	447	1120	543	329
Shifts	386	264	650	48	698	312	397	1095	709	0
Access LUT	0	72	72	12	84	84	108	192	192	0
Absolute	3	3	6	9	15	12	3	18	15	0
Branching	0	1	1	5	6	6	130	136	136	0

TABLE II: Complexity breakdown of proposed rate-control vs. non-perceptual and perceptual (SSIM without scaling)

where $(bg(x, y)$ is background luminance, in this case the higher of the two pixel pair values. These two perceptual models can be combined by first rearranged the SSIM luma function as $1-SSIM_l$, making it in-line with common perceptual principles. However, to consider this a perceptual cost, it should be scaled by the JND background luminance masking visibility threshold [11]. This is shown in equation (10)

$$Luma\,Cost(x, y) = (2^b - 1) \times (1 - SSIM_l)^{JND_l} \quad (10)$$

where b is bit-depth. Combining these two perceptual models in this way allows the SSIM luma function to be more non-linear due to the JND luma function, while the bit-depth range enables a pixel cost to be associated. This like SASD utilises a LUT to retrieve perceptual normalisation cost. Also the proposed luma cost uses the Hadamard transform to eliminate self-symmetry, but the cell distances of 1, 2 or 4, will be downscale by factors of $1/2$, $1/16$ and $1/64$ respectively. As an extension of the initial PAS test, a second stage of filtering is applied using these block sides which involves using the edge detection technique as discussed earlier, but on the respective corners. Since the perceptual normalised process involves Hadamard processing, the bottom right corner of the 8x8 block is always equal to zero and so is not tested.

The first two stages of the proposed perceptual rate-control activity cost work with the sub-block sample, 8x8 block's sides, and they act as an efficient means to eliminates false triggers, only permitting full assessment if it passes a series of thresholds. Table II shows the complexity breakdown of the proposed rate-control verses the existing non-perceptual and perceptual alternative SSIM. Since the proposed rate-control operates on top of the existing rate-control Hadamard (RC Had) function, the respective complexity are aggregated. As illustrated in figure 3, this means that the proposed algorithm is always more complex than the existing rate-control Hadamard function, but far less than SSIM based alternatives. Stages 1 and 2 reflect the two early termination points used to minimise perceptual overhead. While the total number of operations for performing the proposed algorithm is high, through the use of early termination points, stages of 1 and 2, full complexity cost are minimised. For sub-blocks which are no longer of interest the complexity cost can be counted at stages 1 and 2. Comparing the proposed perceptual approach to an existing perceptual technique of SSIM (based on the JM H.264/AVC [13], and assuming constants are pre-calculated) shows the high number of multiply and/or divide for the same 8x8 sub-block. Here, the complexity of SSIM is shown without considering the scaling that would need to occur. Therefore, the proposed perceptual rate-control works at the sub-block level of 8x8 with existing Hadamard based rate-control, offering early termination points whilst being processor-friendly.

C. Modified Decoder

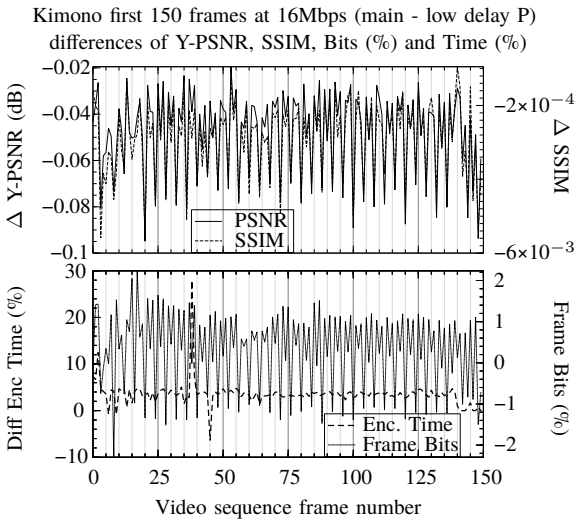
These proposed developments have been possible due to modelling based using empirical methods, but also by adapting the HEVC decoder as a tool to visually analyse their effect. Creating a visual perceptual assessment tool is motivated by the need to analyse how these new and existing techniques behave using real data. This becomes more crucial when visualising the partitioning and quantisation information from the signalling information. This is possible within the decoder by establishing a secondary stream reconstructed with the signalling information superimposed. To indicate the effect of these algorithms, a heatmap is produced, where blue is low and red is high. In certain cases a fixed sub-block assessment size of 8x8 is used to reflect the sub-block sampling approach used in the proposed algorithms.

V. Results and Discussion

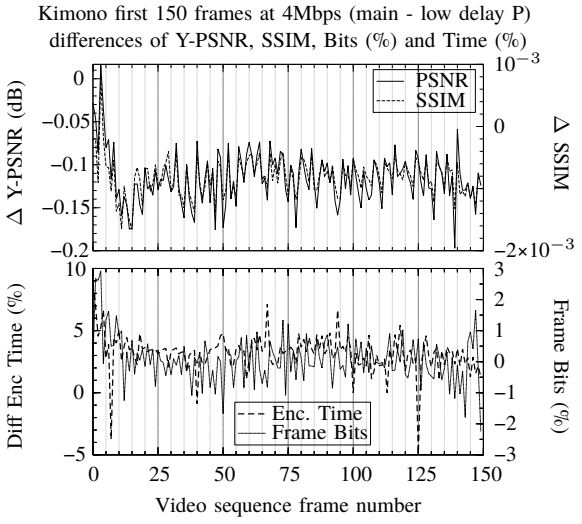
The low complexity perceptual solution proposed is designed to enable PVC in the scenarios illustrated in figure 1. These are aimed at capturing medium and low activity video, therefore, two high definition video sequences (of 1080p resolution) are used. For medium activity, BasketballDrive (50 fps), while for low activity Kimono (24fps), both encoded for 150 frames under the same HEVC main profile at 16Mbps, 4Mbps and 1Mbps. The use of BasketballDrive is to reflect the typical video being encoded of a set of large moving objects on a static background. Conversely, Kimono has a single person in the frame, similar to video calling scenarios, where background and foreground cam be more easily separated.

	Kimono Low Delay P			BasketballDrive Random Access		
Rate (Mbps)	Time (%)	Y-PSNR (dB)	SSIM	Time (%)	Y-PSNR (dB)	SSIM
1	3.06%	-0.1821	-0.0027	2.43%	-0.5298	-0.0110
4	3.12%	-0.1154	-0.0008	2.10%	-0.3353	-0.0047
16	3.51%	-0.0476	-0.0003	3.28%	-0.1045	-0.0012

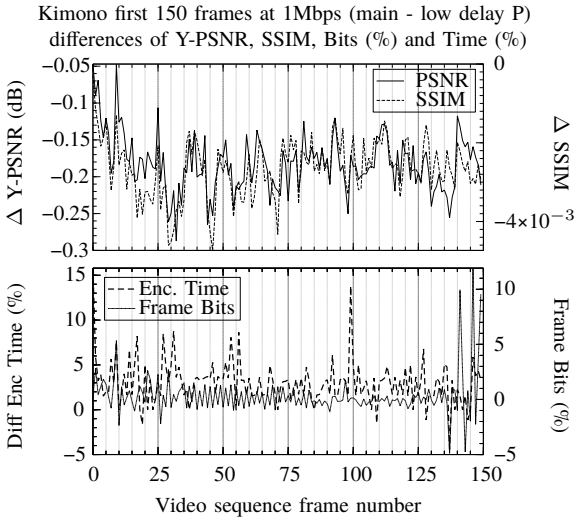
TABLE III: Overall difference by bit-rate for each video



(a) 16Mbps

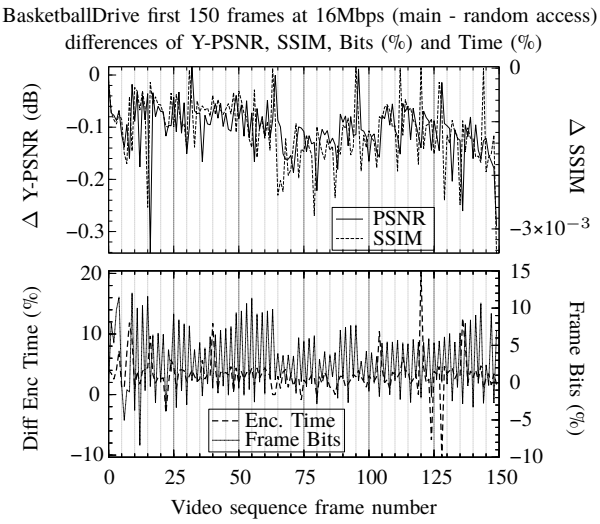


(b) 4Mbps

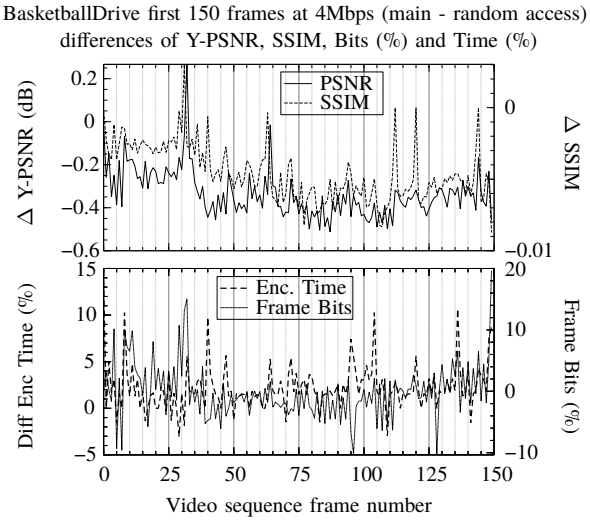


(c) 1Mbps

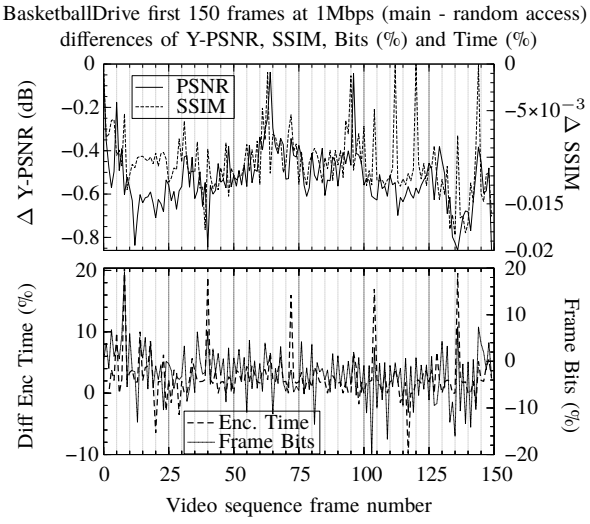
Fig. 4: Kimono first 150 frames



(a) 16Mbps



(b) 4Mbps



(c) 1Mbps

Fig. 5: BasketballDrive first 150 frames

As mentioned in the introduction the scenarios listed in figure 1 can be themed as either storage critical (pre-recorded video) or responsive critical (live interactive video) the two videos have been chosen for those themes. Here, BasketballDrive is selected for storage critical, as shown in figures 1a and 1b, while Kimono for responsive critical, figures 1c and 1d. To match these tests with their scenarios the configuration selected used are random access for storage critical themed scenarios, and low delay P for responsive critical scenarios. Random access is a hierarchical bi-predictive (B-frame) encoding with a group of picture (GOP) of 8 and intra refresh period of 32. Low delay p has a shorter GOP of 4 and uses hierarchical predictive (P-frame) where only the initial frame is an intra frame. In total 12 tests will be conducted altogether, 6 on the original unmodified encoder and 6 on the proposed low complexity perceptual encoder.

The proposed low complexity solution is built upon HEVC version 16 and thus tested against an unmodified HEVC version 16. The encoding timings have been gathered from a system running Ubuntu 15.04 with 7 Gb of RAM and running an Intel Core i7 processor at 2.67 GHz. To enable comparison the differences are reported, either as actual values or as percentages. Table III shows the overall difference of the proposed verses the original by bit-rate for each video sequence, the perceptual losses for both are minor, and the additional timing increases across the encoding sequence of the proposed solution is less than +4%. HEVC encoder produces logs enabling examination on a per frame basis and this is shown in figures 4 and 5, where Y-PSNR, encoding time and frame bits are presented alongside SSIM which is gathered via the modified decoder.

In the results for Kimono in figure 4 the loss in SSIM is insignificant (up to -0.0006, -0.002 and -0.004 for respective bit-rates) compared to the minor loss in PSNR (of -0.1, -0.2 and -0.3 dB respectively). Among the graphs, the changes in picture quality for both PSNR and SSIM tend to follow each other, with only the scale of losses differing. The lower SSIM losses may be attributed to non-perceptually sensitive regions, allowing the average PSNR to be lower, this suggests that the proposed encoder is redistributing bits on perceptual significant distortion. In terms of timing, they are fluctuating around the zero, although, there are instances when the timing differences are shown to dramatically increase or decrease. The extreme high and low in timings across the bit-rates do not cover the same period of frames, so these may be related to balancing content and bandwidth restrictions. In terms of the frame bit usage difference the dynamic is far less, within $\pm 3\%$, except for the least bit-rate of 1Mbps, this could be due to the profile of low delay P which stores additional prediction reference information.

For the results of BasketballDrive in figure 5, the use of random access and medium activity shows the loss in PSNR to be higher than in Kimono, starting from -0.3, then increasing to -0.6 and -0.9 as bit-rates decreases. The perceptual loss shows decreases of -0.003, -0.01 and -0.02 across the same bit-rates but linking set SSIM loss to PSNR loss is difficult due to nature of the video content and calculation of SSIM. Overall, the graphical changes of both PSNR and SSIM are similar, following each other but at times on lower bit-rates, SSIM peaks towards 0, when PSNR reports losses of -0.4 dB. This indicates that PVC can accept losses in PSNR without affecting the overall perceptual integrity of the video sequence. Looking at the frame encoding timings, there are significantly more peaks and troughs in random access compared to low delay P. These may be due to the high compression and intensive searching during motion estimation for B-frames, which is affected by the encoders ability to find J_{min}. For the proposed solution this affects the number of candidates to consider for full perceptual coding which can affect the final selection. As the algorithms are sub-block based, a visual simulation of their effects on an encoded frame may provide some insight of how the perceptual algorithm is behaving.

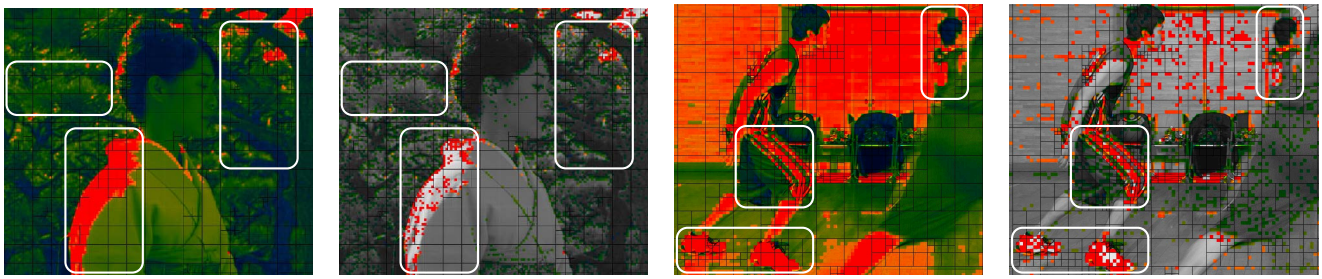
The modified decoder can assist to understand the proposed encoder's behaviour as it shows the effects of signalling in terms of sub-block partitioning, quantisation and by allowing the simulation of different distortion algorithms. Using the encoded bitstreams and extracting a frame from the middle of the sequence, in this case frame 77, the cropped heatmap images under different forms of assessment are shown in figures 6 to 8. Please note rate-control activity for both existing and proposed use the incoming original video sequence frame, so its results will be identical irrespective of bit-rate.

Figure 6 indicates how the proposed rate control which works in addition to the existing rate-control activity assessment places greater activity cost for high intensity regions with texture. This means that during mode decision, the influence of rate-control in bit-budget allocation and the perceptual distortion cost during RDO can lead to larger blocks in perceptually homogenous regions. Figures 7 and 8 show distortion measured in non-perceptual (SSE), perceptual (SSIM), proposed (SASD) and by quantisation (QP), where SSE and SSIM are heatmaps, while SASD and QP refer to heatmaps where triggered. This means that SASD reflects the research design as shown in figure 3 and for a low activity video sequence like Kimono even at 1Mbps there SSE and SSIM scores are low. QP reflects the level of quantisation set for the bits are allocated, otherwise if it is prediction only or block re-use it will remain greyscale.

The perceptually highlighted regions in figure 6b shows increased partitioning within the proposed encoder under figure 7. Similarly, in figure 8, the use of larger block sizes leads to very similar SSIM heatmaps. Overall, the proposed encoder use of larger blocks are on the boundaries between homogeneous and textured regions, suggesting some tolerance is provided under PVC.

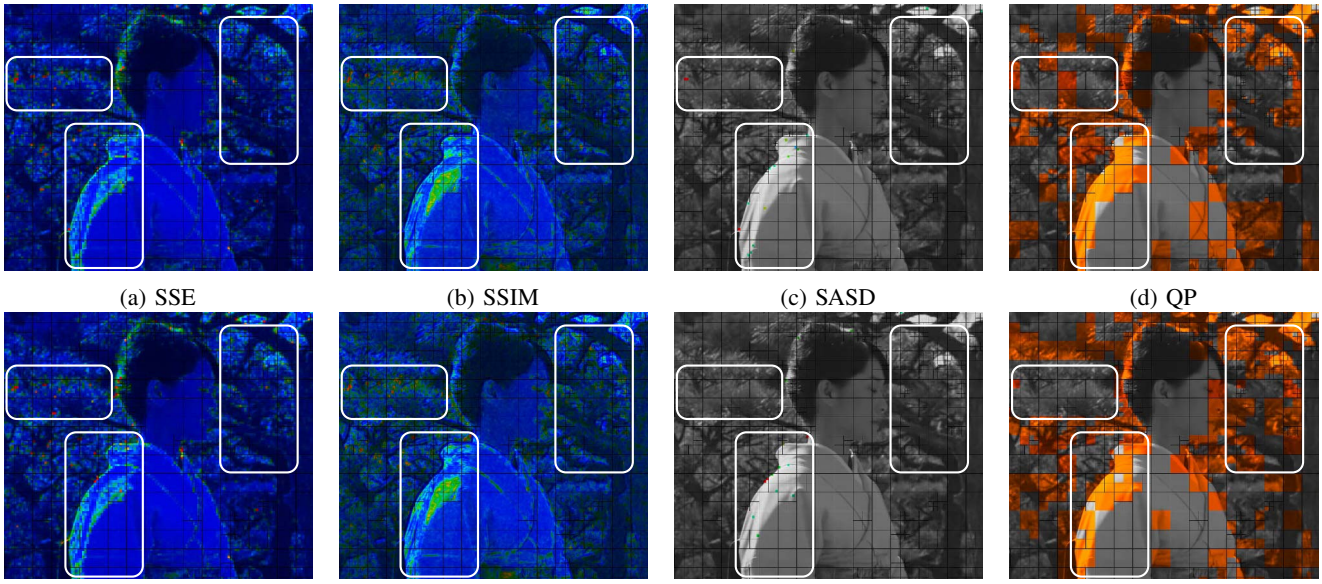
VI. Conclusion and future work

As more video services are accessed on portable devices, using heterogeneous and wireless sensor networks, the need for low complexity perceptual video coding increases. This paper presented, a low complexity perceptual solution capable for assessing candidates individually for both medium and low



(a) Existing (b) Proposed (c) Existing (d) Proposed

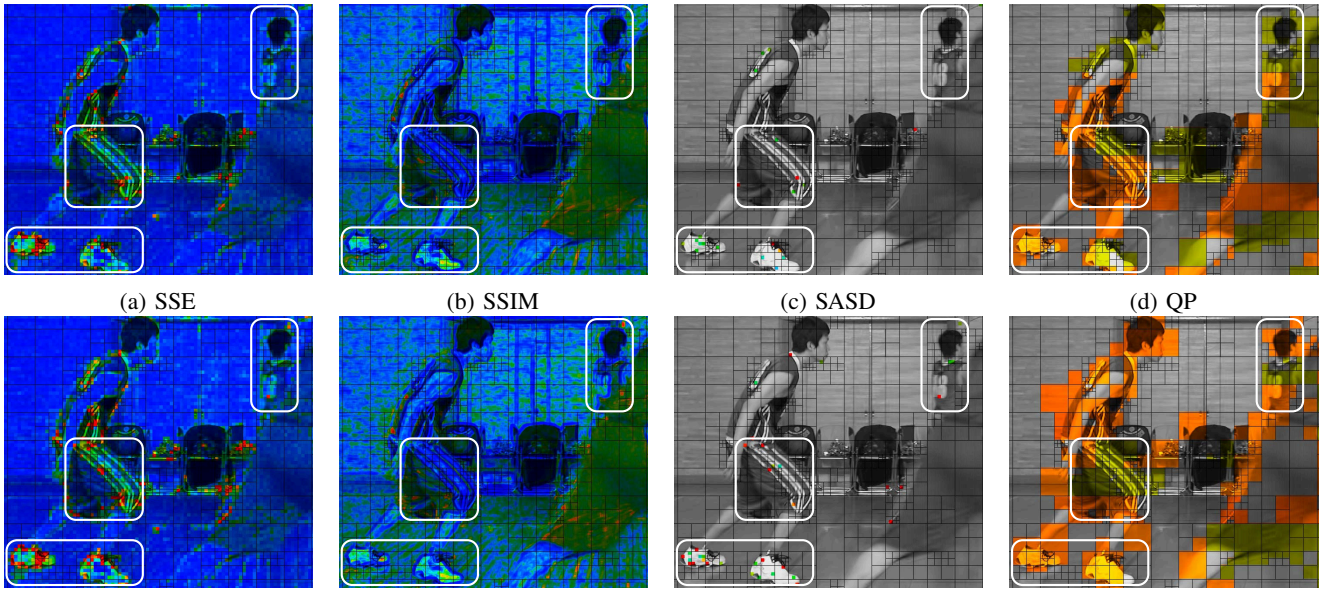
Fig. 6: Rate control activity on Kimono and BasketballDrive - blue is low, red is high



(a) SSE (b) SSIM (c) SASD (d) QP

(e) SSE (f) SSIM (g) SASD (h) QP

Fig. 7: Kimono at 1 Mbps, top row existing, bottom row proposed - blue is low, red is high



(a) SSE (b) SSIM (c) SASD (d) QP

(e) SSE (f) SSIM (g) SASD (h) QP

Fig. 8: BasketballDrive at 4 Mbps, top row existing, bottom row proposed - blue is low, red is high

activity video sequences. This can result in perceptually robust encodings for minimal increases in timing, < +4%. Other perceptual solutions for rate-control and mode-decision adapt existing perceptual algorithms with high levels of complexity and not for access each candidate. This paper presents a perceptual solution without the need for statistical calculations of SSIM or the high complexity operations used to scale SSIM into existing distortion metric space. As such the proposed solution can evaluation each mode-decision and rate-control candidate individually while maintaining a low complexity overhead. This should be extended to the prediction stage, the initial stage where sub-block assessment occurs, however issues of complexity tackled in this paper require greater sensitivity during prediction.

References

[1] A. Ortega and K. Ramchandran. ``Rate-distortion methods for image and video compression". In: *IEEE Signal Processing Magazine* 15.6 (1998), pp. 23–50. ISSN: 1053-5888.

[2] Weisi Lin and C-C Jay Kuo. ``Perceptual Visual Quality Metrics: A Survey". In: *Journal of Visual Communication and Image Representation* 22.4 (2011), pp. 297–312.

[3] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, et al. ``Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric". In: *IEEE Transactions on Circuits and Systems for Video Technology* 20.11 (2010), pp. 1614–1624. ISSN: 1051-8215.

[4] T. Richter. ``A Global Image Fidelity Metric: Visual Distance and its Properties". In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*. 2013, pp. 369–373.

[5] Y.G. Joshi, P. Shah, J. Loo, et al. ``Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6.

[6] Y.G. Joshi, J. Loo, P. Shah, et al. ``A novel low complexity Local Hybrid Pseudo-SSIM-SATD distortion metric towards perceptual rate control". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6.

[7] Hong Ren Wu, Weisi Lin, and King Ngi Ngan. ``Rate-perceptual-distortion optimization (RpDO) based picture coding --- Issues and challenges". In: *Digital Signal Processing (DSP), 2014 19th International Conference on*. 2014, pp. 777–782.

[8] Jens-Rainer Ohm, Gary J. Sullivan, Benjamin Bross, et al. *High Efficiency Video Coding (HEVC)*. Joint Collaborative Team on Video Coding (JCT-VC). 2013.

[9] Zhou Wang, A. C. Bovik, H. R. Sheikh, et al. ``Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[10] A. Horé and D. Ziou. ``Is there a Relationship between Peak-Signal-to-Noise Ratio and Structural Similarity index measure?" In: *IET Image Processing* 7.1 (2013), pp. 12–24.

[11] Chun-Hsien Chou and Yun-Chin Li. ``A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile". In: *IEEE Transactions on Circuits and Systems for Video Technology* 5.6 (1995), pp. 467–476.

[12] Tung-Hsing Wu, Guan-Lin Wu, and Shao-Yi Chien. ``Bio-inspired Perceptual Video Encoding based on H.264/AVC". In: *Proc. IEEE Int. Symp. Circuits and Systems ISCAS 2009*. 2009, pp. 2826–2829.

[13] Karsten Sühring. *H.264/AVC Reference Software JM*. URL: http://iphome.hhi.de/suehring/tml/.