# Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level

Yetish G. Joshi, Purav Shah, Jonathan Loo and Shahedur Rahman
Computer & Communications Engineering Department,
School of Science and Technology, Middlesex University, The Burroughs, London NW4 4BT
{y.joshi, p.shah, j.loo, s.rahman}@mdx.ac.uk

***Abstract*—Within a video encoder the distortion metric performs an Image Quality Assessment (IQA). However, to exploit perceptual redundancy to lower the convex hull of the Rate-Distortion (R-D) curve, a Perceptual Distortion Metric (PDM) modelling of the Human Visual System (HVS) should be used. Since block-based video encoders like H.264/AVC operate at the Sub-Macroblock (Sub-MB) level, there exists a need to produce a locally operating PDM. A locally operating PDM must meet the requirements of Standard Traditional Distortion Metrics (STDMs), in that it must satisfy the Triangle Equality Rule ($\lhd$). Hence, this paper presents a review of STDMs of SSE, SAD and SATD against the perceptual IQA of Structural Similarity (SSIM) at the Sub-MB level. Furthermore, this paper illustrates the Universal Bounded Region (UBR) by block size that supports the triangle equality rule ($\lhd$) within the Sub-MB level, between SSIM and STDMs like SATD at the prediction stage.**

## I. Introduction

Encoders such as MPEG4/AVC (H.264/AVC) and more recently H.265 are deemed as block-based video encoders [1], [2], as they select a prediction mode for a given block with the minimum of pixel difference - residue. This is extended when inter coding is considered as the grouping of (Sub) Macroblocks (Sub-MB) with the minimum amount of motion vectors for the least amount of distortion. Therefore, majority of the block can be represented with signalling and quantisation of the residual pixel difference. This is represented by the Rate-Distortion (R-D) curve in equation (1) in [3], where the effects of lambda ($\lambda$) to maintain a given bit rate ($R$) as part of Rate Control must be assessed by the distortion metric ($D$).

$$J_{min\,energy} = \lambda_{quant} \times R_{bit\,rate} + D_{dist\,metric} \qquad (1)$$

Hence, the quantisation benefit of lowering the bit rate must be factored with any cost increase in the distortion measured, leading to the search for $J_{min\,energy}$, which can be considered to be an optimum point of operation for the encoder. In particular, the role of the distortion metric is significant within the R-D curve, described in [4] as a convex hull. In the context of the front-end of the encoder, the distortion metric is used in three main areas; selection of the prediction modes, choosing various block sizes during mode decision and assessing the level of activity for the incoming MB when taking Rate Control into account. This process can be illustrated in figure 4 as stages ‘1’ (Distortion Metric - red box), ‘2’ (Mode Decision - green box) and ‘3’ (Rate Control based on [5] - blue dashed outline) respectively.

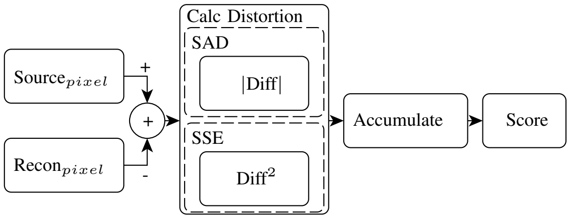


Fig. 1. Distortion Metric Operation with SAD or SSE IQA

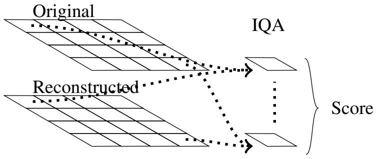


Fig. 2. Standard Traditional Distortion Metric (STDM)’s based IQA

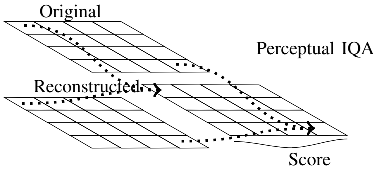


Fig. 3. Perceptual IQA

However, it was discussed in [4], [6] that distortion metrics within video encoders should ideally be based upon the Human Visual System (HVS), though due to reasons of complexity and lack of tractable scoring HVS solutions, they have not been integrated at the block-base level of a video encoder. Instead, Standard Traditional Distortion Metrics (STDMs), such as Sum of Square Errors (SSE) and Sum of Absolute Difference (SAD) are used at the block-base level. These STDMs are simple to operate and tractable, where every pixel difference is uniformly accumulated towards an overall distortion score as shown in figure 1.

An advantage of modelling the HVS is perceptual sensitivity

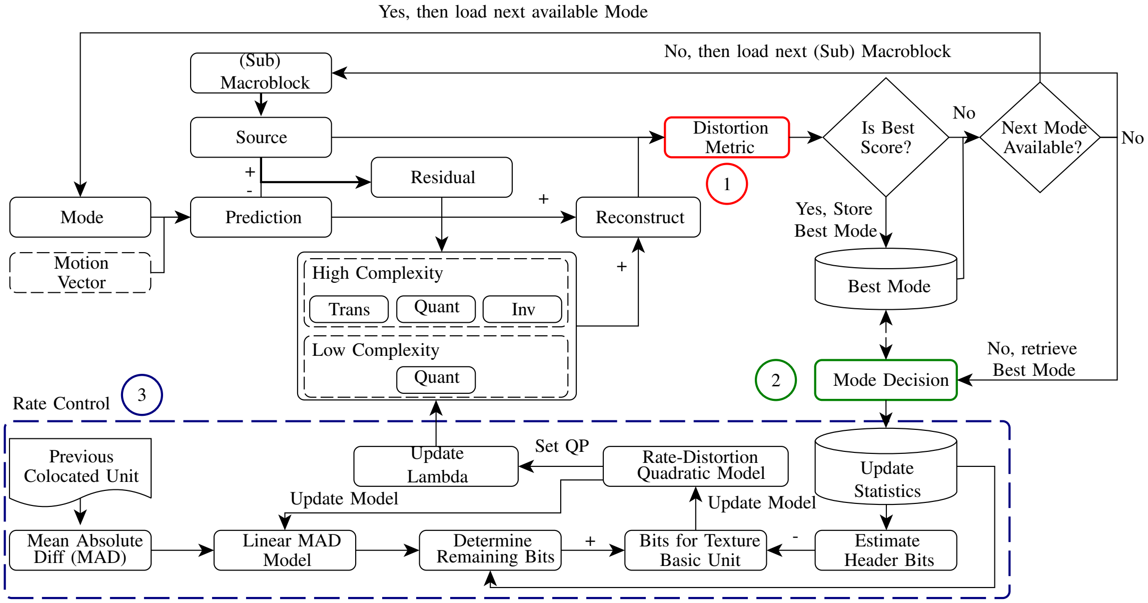


Fig. 4. Front-End Block Based Video Encoder System Level Overview with Rate Control

of relative lighting conditions and structural information can be evaluated compared to absolute pixel difference accumulation of STDMs. This is illustrated in figures 2 and 3 and reinforced in [7]. In terms of a perceptual based modelling, the convex hull can be closer to the origin as distortion is non-uniformly weighted like Just Noticeable Distortion (JND) [8] and Contrast Sensitivity Function (CSF) [9].

JND and CSF perceptual based models reflect the nature of HVS's sensitivity to varying lighting conditions. In these models, least sensitivity is applied to darker regions where objects or texture can be less distinguishable [7]. When edges are visible to the HVS, cognitive sensitivity allows for objects to be recognised by their structural information [10]. Therefore, the HVS relies upon structural information based upon relative lighting conditions to recognise and track objects. Compared to STDMs, where equal weighting is provided, perceptual Image Quality Assessment (IQA) identifies perceptual clues worth retaining and perceptual redundancy that can be exploited for better bit budget utilisation.

However, while these early models of JND and CSF showed the promise to distinguish on perceptual terms, they are complex to implement and operate at the frame level. This computational burden of early models motivated a second generation of application specific perceptual models, primarily for perceptual based coding in video-calling application [11] [12]. Here they focused on reducing the computational complexity by simplifying aspects of these perceptual based models and combining other perceptual based models such as edge detection to produce a multi-HVS perceptual model.

While neither of these application based models replaced the distortion metric, they highlighted the need to do so. The current third generation of HVS modelling took the initiative to consider this direction of modifying the distortion metric and replace it with a multi-HVS based model. This was attempted in [13], [14], however, a perceptual-based model faces the challenge of being implemented within the encoder workflow as a low processing envelope as well as operating as a locally independent operations as discussed in [6]. This has not been successfully achieved to date.

## II. A Way Towards Perceptual IQA - Structured Similarity (SSIM)

Structural Similarity (SSIM) [15], a low complexity perceptual Image Quality Assessment (IQA) that takes into account the structural information based on relative lighting conditions and is described in equation (2),

$$SSIM(org,rec)=\frac{(2\mu_{org}\mu_{rec}+C_1)\times(2\sigma_{org,rec}+C_2)}{(\mu_{org}^2+\mu_{rec}^2+C_1)\times(\sigma_{org}^2+\sigma_{rec}^2+C_2)} \quad (2)$$

where, $\mu_{org}$ and $\mu_{rec}$ represent the mean of the original image block and reconstructed image block, $\sigma_{org}^2$and $\sigma_{rec}^2$ are the standard deviations respectively, $\sigma_{org,rec}$ is the covariance, $C_1$ and $C_2$ are constants which are calculated based upon the bit depth to stabilise the equation. An extensive study discussed in [16] showed a range of perceptual based IQA's available (including variations of SSIM) being tested and it was concluded that SSIM performed well whilst offering a low processing overhead.

Compared to STDM, SSIM does not support the Triangle Equality Rule ($\lhd$) natively. In terms of the video encoding, the triangle equality rule is where an image triplet of original, predicted and difference are considered; the distortion score of each pair should be such that the distortion score of one should equate to the summation of the other two sides [17].

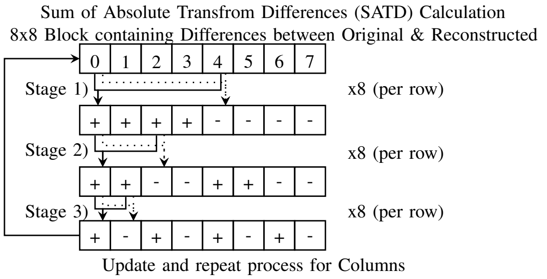

Sum of Absolute Transfrom Differences (SATD) Calculation
8x8 Block containing Differences between Original & Reconstructed
0 1 2 3 4 5 6 7
Stage 1)
x8 (per row)
+ + + + - - - -
Stage 2)
x8 (per row)
+ + - - + + - -
Stage 3)
x8 (per row)
+ - + - + - + -
Update and repeat process for Columns

Fig. 5. SATD Operational Block Diagram.

In [13], SSIM has been scaled using logarithmic functions within the distortion metric space of MSE in order for SSIM to support the triangle equality rule ($\trianglelefteq$) and this approach only works at the Group of Picture (GOP) level and being unable to adapt to the local changes. Furthermore, in [13], a perceptual measure is scaled into the distortion space of MSE, where the distortion scale can potentially be wide. Typically, for an 8-bit Luma pixel depth this means a theoretical maximum of $255^2$.

The applications of perceptual IQA limited to GOP and frame level, restricts the effect on bringing R-D curve closer to the origin. Therefore, the process of mapping perceptual IQA on non-Perceptual Distortion Metric (PDM) should be extended to the Sub-MB level to reflect the operation of a block based video encoder. In addition, non-PDM should be evaluated in terms of their complexity and potential range of distortion scores for a low processing overhead and limited range of values respectively.

Thus, this paper will investigate whether a SSIM based PDM can exist at the Sub-MB level. This will be done by assessing Sub-MBs simultaneously under SSIM and STDM, evaluating whether SSIM operates within a closed distortion metric space of STDMs. A closed distortion metric space will indicate that SSIM can be scaled to satisfy the triangle equality rule ($\trianglelefteq$). Hence, a future block-based encoder using a scaled-SSIM-PDM can achieve a lower convex hull R-D curve.

## III. SSIM within the Distortion Metric Space of STDMs

The independent dimensionless pixel level evaluation of STDMs such as SSE and SAD are scalable, unaffected by adjacent pixel differences. Therefore, STDMs are unable to appreciate the significance of inherent visual clues like structure or texture within a Sub-MB. Another distortion metric to consider is SATD, which utilises the Hadamard Transform and is designed to be processor friendly as shown in figure 5, already used in H.264/AVC's back-end [18].

Unlike other STDMs that have a high dependency on computational loops, SATD is an efficient alternative [18] as it utilises shifts, addition and subtraction. However, these STDMs are weightless metric, meaning each difference is treated equally. In perceptual IQA, it has not yet been discussed whether the amount of neighbouring pixels would affect the performance at the Sub-MB level. It is important to understand that SSIM is an averaging of a series of sliding windows of local SSIM's and hence, it is about determining the size of the block and the amount of overlap between blocks [19]. It was shown in [15] that a block size of 8x8 pixels is recommended to provide a stable result and a greater degree of overlap between blocks would provide an accurate SSIM result. SSIM also supports the small 4x4 block size of Sub-MB [20].

Having a PDM will influence both intra and inter blocks across each of the stages as shown in figure 4 as part of the Perceptual Framework design. To minimise the processing load, the SSIM window size will be equal to the Sub-MB size. This approach can be extended if needed, to produced a more accurate SSIM, by using smaller window sizes and overlapping windows at the expense of additional processing.

## IV. An Investigation into Perceptual IQA at the Prediction Stage with Sub-Macroblocks

In order to lower the convex hull of the R-D curve, it is necessary to have a PDM working at the Sub-MB level. This paper introduces SSIM at the prediction level in order to assess its feasibility to work at the Sub-MB level against STDMs. The results presented in figure 6 and figure 7 have been extracted from the JM18.4 H.264/AVC [18] encoder, which has been modified to incorporate SSIM at the prediction stage with SSIM window size equal to the Sub-MB size. The default configuration file for JM18.4 complies predominately with the recommendations set in [21] with only minor changes required.

The video sequence used for these tests is chosen as the Foreman video, with QCIF resolution of $176 \times 144$ pixels and consists of three frames. At the prediction stage in the Sub-MB level, the iterative operations result in 900k and 700k samples captured for the 4x4 and 8x8 block respectively across both the inter frames in the test video sequence.

The test results in Figure 8 were obtained using higher CIF resolution based video sequences of varying content with only 4x4 and 8x8 inter blocks considered to validate the earlier findings of figure 7.

Focusing on the Intra graphs as shown in figure 6, a concentration of samples can be described close to the origin highlighting the statistical perceptual similarity of intra prediction. In 4x4, a broad range of 1-SSIM values exist for a limited range of STDM score, suggesting that SSIM evaluates with greater sensitivity when in smaller block sizes.

In terms of the 8x8 Intra graphs, the results seem more narrow, usually with most samples concentrated within the first 0.25 of (1-SSIM) range. This suggests that 8x8 does encounter predictions that are favourable for SSIM than STDMs.

The results for the Inter blocks as shown in figure 7 have improved definition of the distortion metric space than of Intra. This is because RDO is enabled leading to permutations of mode predictions and motion vector predictions being considered. As such, 4x4 blocks of Inter extend 1-SSIM to 0.75, where as in the 8x8 configuration, the shape of the distortion metric space is beginning to appear with trails of samples extending along the x-axis beyond 1. The samples where (1-SSIM) is $<1$, more perceptual information is stored, conversely; samples $>1$ have high amounts of blocking artefacts making it perceptually unrecognisable.

Upon analysing the distortion score ranges, it is found that SSE has the highest range of 125k for 4x4 and 250k for 8x8. Theoretically, this could be as high as 1 million and 4 million respectively in this case which is highly unlikely. With regards to the scales recorded against the theoretical highs, this represents as a fraction 1/8th of 4x4 and 1/16th of 8x8 SSE's distortion metric space. For SAD and SATD, they cover a greater proportion of the maximum possible scores, $\approx 3/10$ and $1/4$ for 4x4 and 8x8 respectively. Knowing that SAD and SATD have a smaller dynamic range and the samples cover a larger proportion of the distortion metric space, SATD allows for any potential model to be mapped with greater coverage.

Overall, analysing figures 6 and 7 by block size, indicates that two scaled-SSIM models by block size are required to produce a scaled-SSIM-PDM, as the graphs illustrate Intra to be a limited version of the Inter.

As SATD is the preferred distortion metric in [18] due to its processor friendly operations, and in order to validate the relationship of 1-SSIM and STDM, further results were gathered. Following the results presented in figures 6 and 7, it has been possible to replicate the relationship of perceptual IQA vs. non-PDM using SSIM and SATD respectively with higher resolution video sequences. This is shown in figure 8, where CIF resolution is used with the number of samples gathered in excess of four million. The overall shape is the same as seen earlier with Foreman (QCIF), though depending on the nature of the video the scores differ. This shows that a scaled-SSIM-PDM can exist within the STDM distortion metric space that satisfies the triangle equality rule ($\trianglelefteq$).

Therefore, these findings are significant as it reflects that a Universal Bounded Region (UBR) by block size at the Sub-MB level exists, irrespective of video resolution or the type of video sequence. This supports the case for the SSIM is mapped against an STDM space.

## V. Conclusions and Future Work

HVS offers the ability to assess perceptually significant and redundant information. To effectively implement a perceptual based HVS model at the encoder system level, it requires to be integrated at the Sub-Macroblock level. However, the triangle equality rule ($\trianglelefteq$) inhibits perceptual IQA such as SSIM from being adopted at the Sub-MB level. This paper has presented the evidence that a Perceptual IQA - SSIM, at a Sub-MB level has a relationship with STDMs. This was further confirmed by higher resolution video, illustrating that this relationship is independent of the video resolution and type of sequence. Hence, a Perceptual Distortion Metric (PDM) can be modelled by scaling SSIM within what is labelled as the Universal Bounded Region (UBR) by block size, thus satisfying the triangle equality rule ($\trianglelefteq$).

Furthermore, a Perceptual Framework can be designed around PDM to affect the highlighted regions of distortion metric, mode decision and rate-control as shown in figure 4.

## References

[1] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649 –1668, dec. 2012.

[3] H. Everett III, "Generalised Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," *Operations Research*, vol. 11, no. 3, pp. 399 – 417, 1963.

[4] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[5] Z. Li, W. Gao, F. Pan, S. Ma, K. Lim, G. Feng, X. Lin, S. Rahardja, H. Lu, and Y. Lu, "Adaptive Rate Control for H.264," *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 376 – 406, 2006.

[6] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[7] H. . R. . Wu and K. . R. . Rao, Eds., *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2005.

[8] C.-H. Chou and Y.-C. Li, "A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 467–476, 1995.

[9] J. Yogeshwar and R. J. Mammone, "A New Perceptual Model for Video Sequence Encoding," in *Proc. Conf. th Int Pattern Recognition*, 1990, pp. 188–193.

[10] S. J. S. Robin E. N. Horne, Ed., *The Colour Image Processing Handbook*. Springer Berlin / Heidelberg, 1998.

[11] R. Jin and J. Chen, "The Coding Rate Control of Consistent Perceptual Video Quality in H.264 ROI," in *Proc. Int. Symp. Computer Network and Multimedia Technology CNMT 2009*, 2009, pp. 1–4.

[12] X. K. Yang, W. S. Lin, Z. K. Lu, X. Lin, S. Rahardja, E. P. Ong, and S. S. Yao, "Local Visual Perceptual Clues and its use in Videophone Rate Control," in *Proc. Int. Symp. Circuits and Systems ISCAS '04*, vol. 3, 2004.

[13] Y.-H. Huang, T.-S. Ou, P.-Y. Su, and H. Chen, "Perceptual Rate-Distortion Optimization using Structural Similarity Index as Quality Metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1614 –1624, 2010.

[14] A. Bhat, I. Richardson, and S. Kannangara, "A New Perceptual Quality Metric for Compressed Video Based on Mean Squared Error," *Signal Processing: Image Communication*, vol. 25, no. 8, pp. 588–596, 2010.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] W. Lin and C.-C. J. Kuo, "Perceptual Visual Quality Metrics: A Survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297 – 312, 2011.

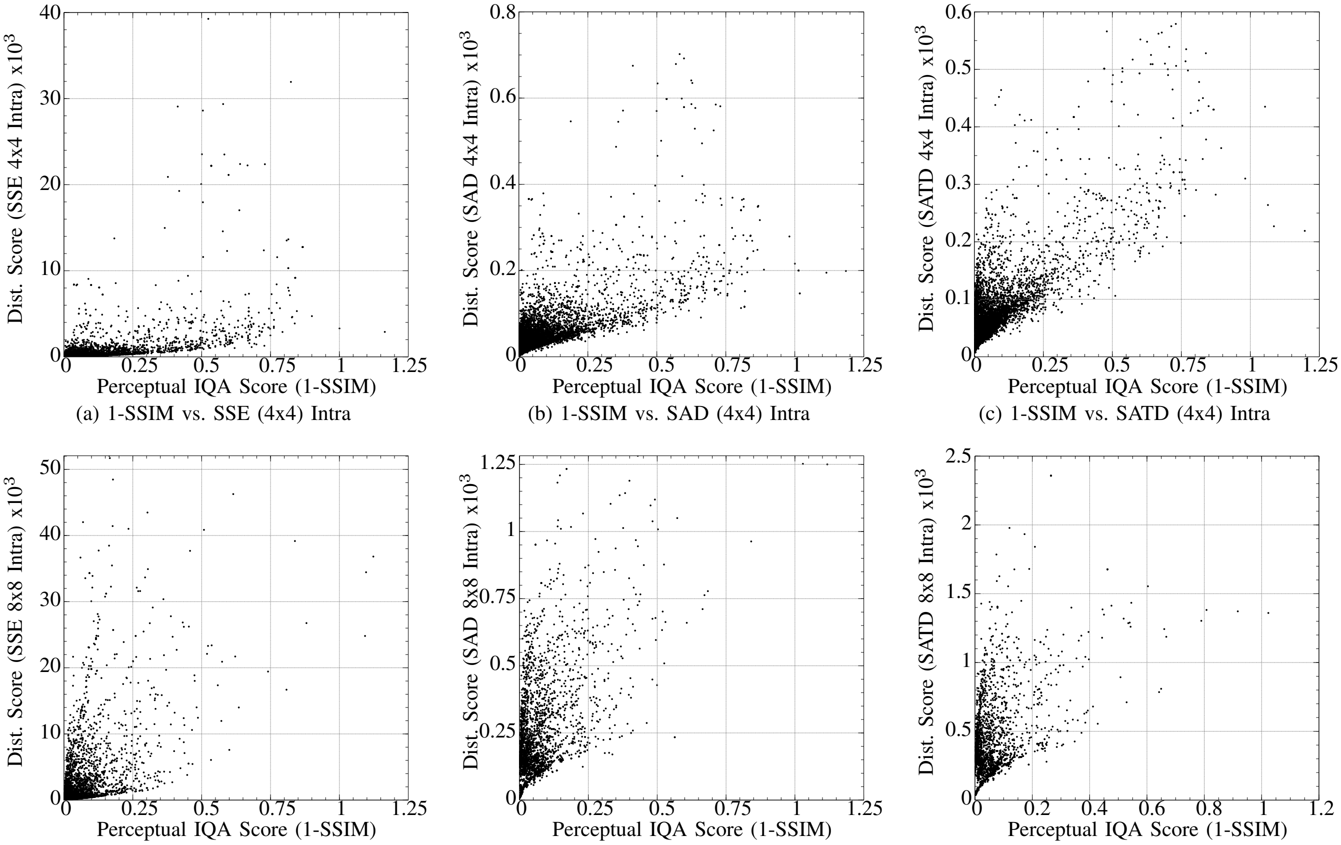
[17] T. Richter, "SSIM as Global Quality Metric: A Differential Geometry View," in *Proc. Third Int Quality of Multimedia Experience (QoMEX) Workshop*, 2011, pp. 189–194.

[18] K. Sühring. H.264/AVC Reference Software JM. [Online]. Available: http://iphome.hhi.de/suehring/tml/

[19] D. Brunet, "A Study of the Structural Similarity Image Quality Measure with Applications to Image Processing," Ph.D. dissertation, University of Waterloo, Sept 2012. [Online]. Available: http://hdl.handle.net/10012/6982

[20] A. Brooks, X. Zhao, and T. Pappas, "Structural Similarity Quality Metrics in a Coding Context: Exploring the Space of Realistic Distortions," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1261 –1273, aug. 2008.

[21] S. G. Tan T.K. and W. T., "(VCEG-AJ10r1) Recommended Simulation Common Conditions for Coding Efficiency Experiments Revision 4." ITU-T SC16/Q6, 36th VCEG Meeting, San Diego, USA, 8th - 10th Oct., 2008, 2008.



(a) 1-SSIM vs. SSE (4x4) Intra

(b) 1-SSIM vs. SAD (4x4) Intra

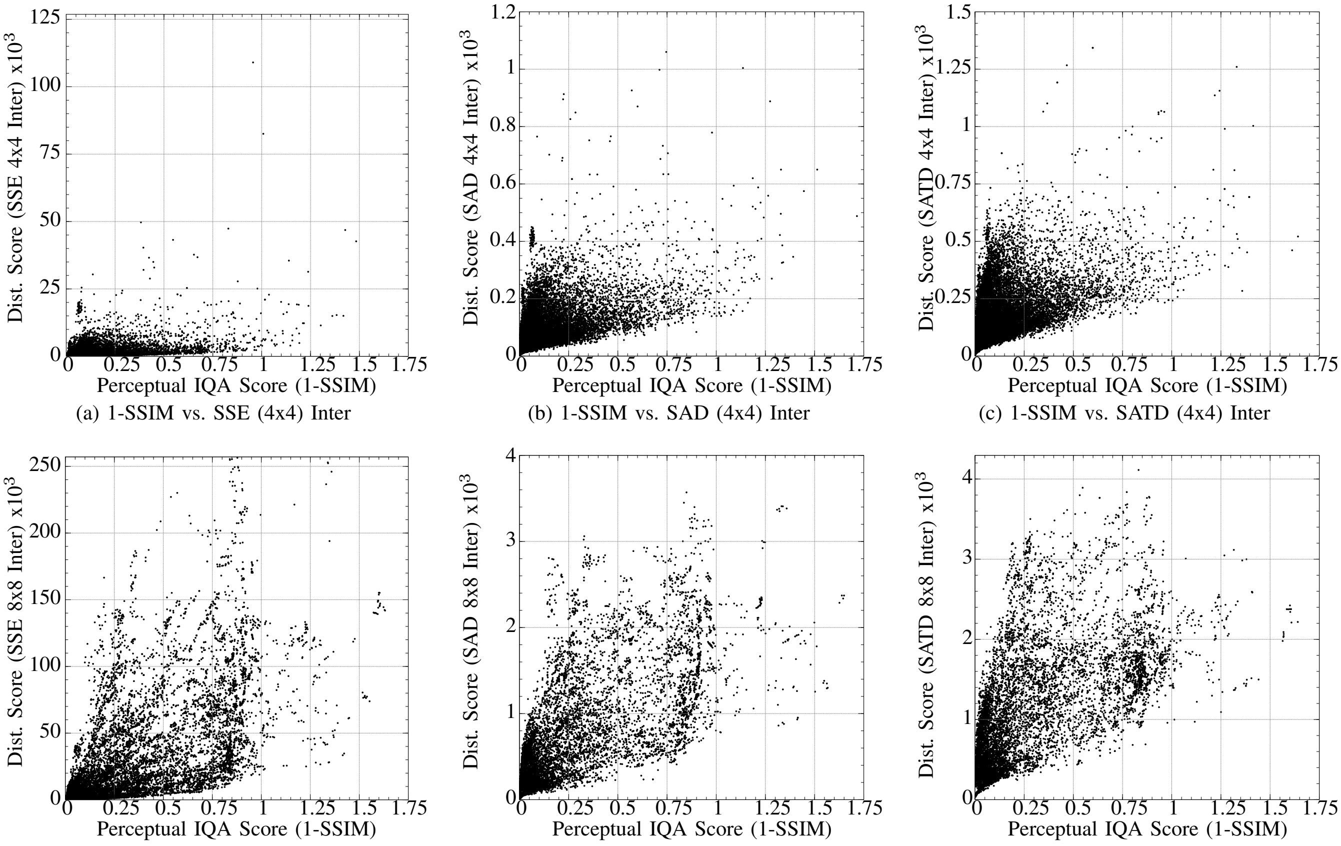
(c) 1-SSIM vs. SATD (4x4) Intra

(d) 1-SSIM vs. SSE (8x8) Intra

(e) 1-SSIM vs. SAD (8x8) Intra

(f) 1-SSIM vs. SATD (8x8) Intra

Fig. 6. Perceptual Image Quality Assessment (IQA) vs. Distortion Metric from 4x4 and 8x8 Intra Blocks. Structural Similarity (SSIM), plotted against Sum of Square Errors (SSE), Sum of Absolute Difference (SAD) and Sum of Absolute Transform Difference (SATD).



(a) 1-SSIM vs. SSE (4x4) Inter

(b) 1-SSIM vs. SAD (4x4) Inter

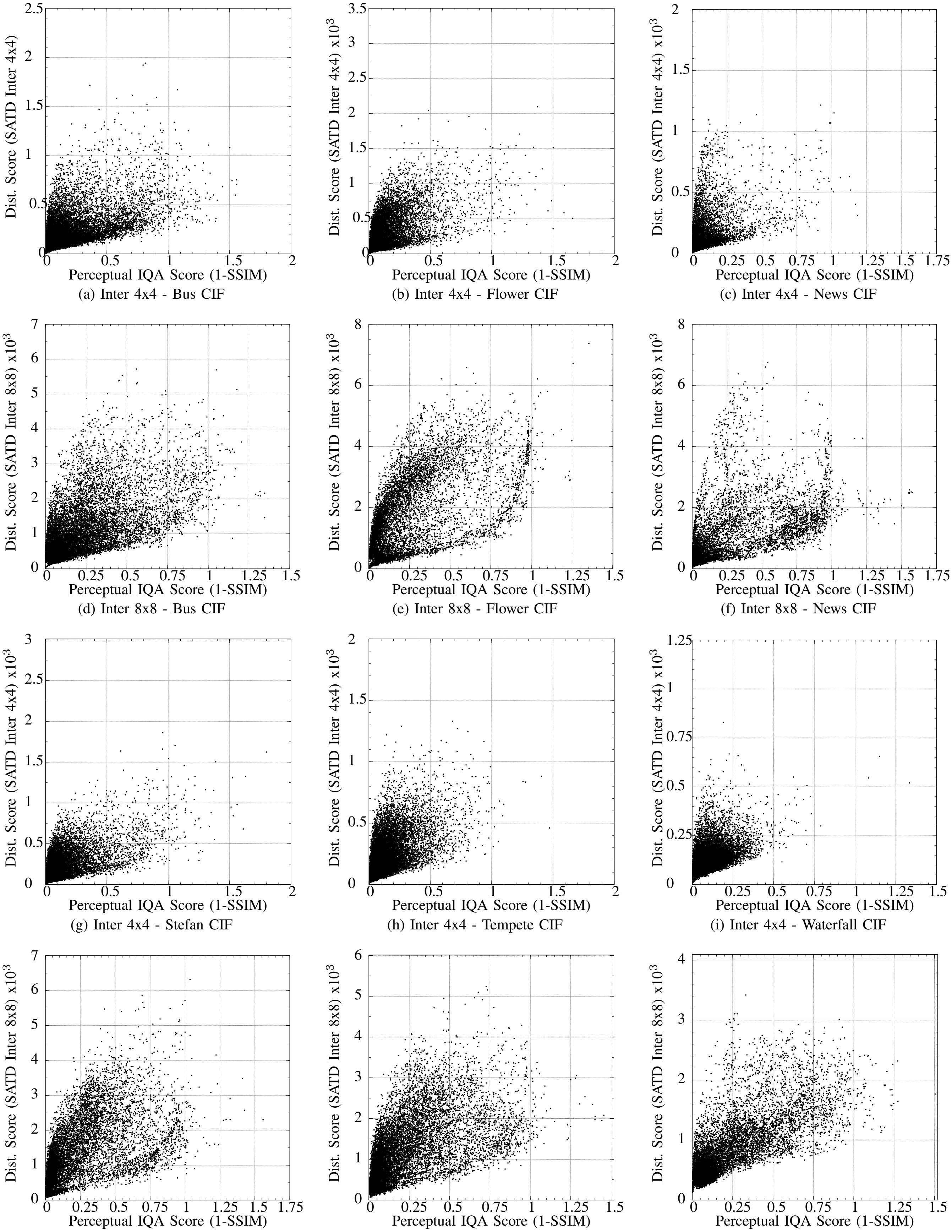
(c) 1-SSIM vs. SATD (4x4) Inter

(d) 1-SSIM vs. SSE (8x8) Inter

(e) 1-SSIM vs. SAD (8x8) Inter

(f) 1-SSIM vs. SATD (8x8) Inter

Fig. 7. Perceptual Image Quality Assessment (IQA) vs. Distortion Metric from 4x4 and 8x8 Inter Blocks. Structural Similarity (SSIM), plotted against Sum of Square Errors (SSE), Sum of Absolute Difference (SAD) and Sum of Absolute Transform Difference (SATD).



(a) Inter 4x4 - Bus CIF

(b) Inter 4x4 - Flower CIF

(c) Inter 4x4 - News CIF

(d) Inter 8x8 - Bus CIF

(e) Inter 8x8 - Flower CIF

(f) Inter 8x8 - News CIF

(g) Inter 4x4 - Stefan CIF

(h) Inter 4x4 - Tempete CIF

(i) Inter 4x4 - Waterfall CIF

(j) Inter 8x8 - Stefan CIF

(k) Inter 8x8 - Tempete CIF

(l) Inter 8x8 - Waterfall CIF

Fig. 8. Perceptual Image Quality Assessment (IQA) vs. Distortion Metric from 4x4 and 8x8 Inter Blocks. Structural Similarity (SSIM), plotted against Sum of Absolute Transform Difference (SATD) for CIF Video Sequences (Bus, Flower, News, Stefan, Tempete and Waterfall). First three frames were used and over four million samples gathered. Graphs show thinned results by a factor of 250.