A study of subjective video quality at various spatial resolutions

Alex Mackin, Mariana Afonso, Fan Zhang, and David Bull

Abstract

In this paper we present the BVI-SR video database, which contains 24 unique video sequences at a range of spatial resolutions up to UHD-1 (3840p). These sequences were used as the basis for a large-scale subjective experiment exploring the relationship between visual quality and spatial resolution when using three distinct spatial adaptation filters (including a CNN-based super-resolution method). The results demonstrate that while spatial resolution has a significant impact on mean opinion scores (MOS), no significant reduction in visual quality between UHD-1 and HD resolutions for the super-resolution method is reported. A selection of image quality metrics were benchmarked on the subjective evaluations, and analysis indicates that VIF offers the best performance.

Index Terms

Spatial resolution, visual quality, 4K

I. Introduction

The video parameter space is being extended to satisfy the desire by consumers for more immersive and high quality video experiences [1]. The unavoidable consequence of increased spatial resolution, frame rate and bit-depth (dynamic range) is higher data rates - a key problem for content providers who want to entertain a wide range of audiences. This issue is currently addressed by increasing compression ratios, through coarser quantisation and/or increased encoding complexity.

While the use of a fixed parameter space within the spheres of broadcast (traditional and IPTV) and cinema has worked well historically due to technological limitations, it has culminated in perceptually redundant information being transmitted for all but the most challenging content - especially as we start to reach the limits of human perception [2] with S/UHDTV video formats [1]. This is because the utility of any video parameter e.g. spatial resolution, bit-depth or frame rate, is inextricably content, display and viewing environment dependent [3]–[9]. Therefore given a sensible baseline, video parameters should be selected in a systemic way, as to provide perceptually optimised experiences.

Before spatial adaptive video formats become commonplace, further research is required to characterise the relationship between visual quality and spatial resolution, such that robust quality metrics can be tested/developed.

In this paper we present the BVI-SR video database, which is used as the basis for a large-scale subjective experiment to investigate visual quality across a range of spatial resolutions and adaptation filters - including a state-of-the-art CNN-based super-resolution method (VDSR) [10]. We then test a selection of quality metrics on the subsequent ground truth to determine their suitability for resolution adaptation.

The remainder of this paper is organised as follows: Section II summarises the state-of-the-art; Section III describes the source sequences and the methodology of the subjective experiment, before results and further analysis is presented Section IV. Conclusions are then provided in Section V.

II. Related Work

Li *et al.* [3] explored seven up-scaling filters (+ reference) on viewer preference scores using a small video dataset (8 source sequences). Their results indicate that the then state-of-the-art filters could not achieve similar perceptual quality to the UHD-1 reference resolution when up-sampling from 1080p (HD) and 720p resolutions. Using a similar test methodology [13], Van Wallendael *et al.* [7] compared UHD-1 and HD resolutions on a larger video dataset (31 source sequences), and found a positive sharpness difference between the reference UHD-1 and

The authors acknowledge EPSRC for funding (grant #: EP/M000885/1).

Fig. 1: A sample frame from the 24 source sequences, which were captured by ourselves, or taken from the Harmonic* [11] and Netflix† [12] video databases.

up-scaled HD resolutions with a Lanczos-3 filter (although content dependence was reported). Both studies used a paired comparison methodology [14], and while useful for discriminating small distortions, it is unrepresentative of typical viewing environments (a high quality reference is generally not available). We address this with a single stimulus methodology [15] and a large number of participants (to mitigate the associated reduction in statistical power).

Recently proposed deep learning super-resolution methods [10] have resulted in major increases in both objective [16] and subjective [17] quality compared to traditional kernel based methods e.g. bicubic/Lanczos-3. However little research has been conducted into their suitability for applications related to spatial resolution adaptation.

The role of video compression is intentionally ignored here, and as even though it is an important facet [18], [19], it increases the number of independent variables that need to be explored e.g. bitrate, QP. The BVI-SR video database will though be used as a platform for further research in this area.

III. Experimental Setup and Methodology

A. Source Sequences

The Bristol Vision Institute Spatial Resolution (BVI-SR) video database[1] (see Fig. 1) contains 24 unique five-second video sequences at 3840×2160 (UHD-1) spatial resolution, 60 fps and 10 bits per colour channel. Half of the sequences were captured natively at 120 fps with a fully open shutter (360°) by ourselves, and were down-sampled by frame averaging to 60 fps [5]. They were then graded in BT.2020 colour space [1] using REDCINE-X software. A further 12 video sequences were selected from the Harmonic [11] and Netflix [12] video databases, and were chosen as the sequences which resulted in the highest uniformity of coverage over Spatial Information [20] (an estimate of high frequency energy in the scene) when including all 24 sequences.

B. Test Sequences

The 24 source sequences shown in Fig. 1 were spatially down-sampled by factors of 2 (1920×1080), 4 (960×540) and 8 (480×270) using two adaptation filters: nearest neighbour (used to simulate the lower resolution as if it was native), and bicubic (commonly used in similar applications). Further adaptation filters [3] such as Lanczos-3 were omitted to reduce redundancy and limit the length of the experiment.

The same method for down-sampling was then used to up-sample the content to the original UHD-1 resolution. A further test condition involved using the super-resolution method VDSR [10] to up-sample the sequences that

[1]BVI-SR will be made available online if the paper is accepted.

had been down-sampled using the bicubic filter (as this was the methodology used for training the CNN) to UHD-1 resolution.

Fig. 2 shows a zoomed region of a frame in the Venice sequence after being down-sampled to 960×540 resolution and then up-sampled to the original UHD-1 resolution. A comparison between the three adaptation schemes and the original UHD-1 frame (bottom-right) shows that: nearest neighbour (top-left) exhibits distinctive blocking artefacts; bicubic (top-right) is over-smoothed; whereas the super-resolution filter (bottom-left) retains most of the spatial detail of the original reference frame. The shape/structure of objects can sometimes be distorted when using CNN-based methods such as VDSR, and is because the filters have no symmetry constraint.

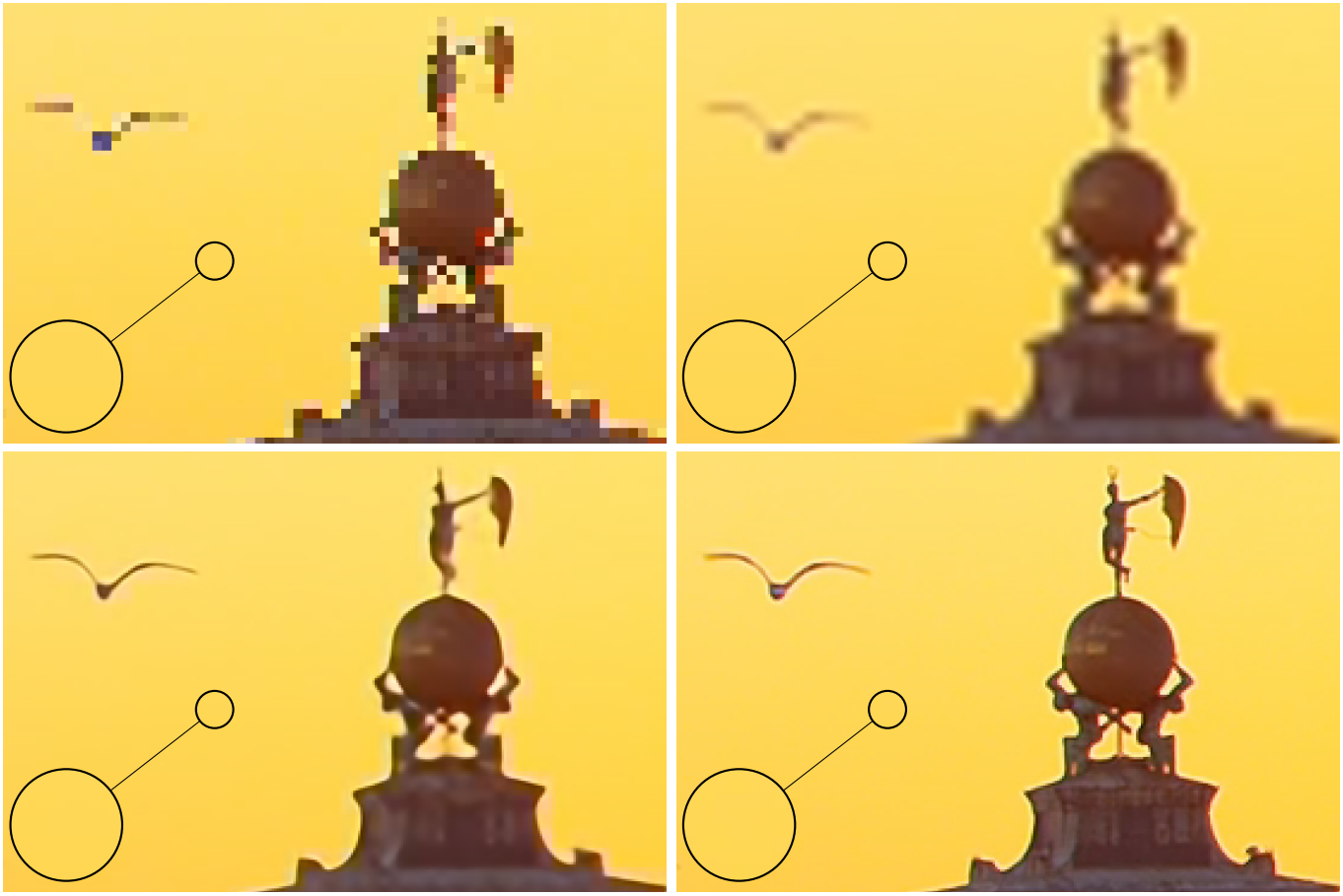

Fig. 2: A zoomed region (4×) from the Venice sequence, comparing the (top-left) nearest neighbour, (top-right) bicubic and (bottom-left) super-resolution filters when up-sampling from 540p to UHD-1 resolution (bottom-right).

The recommended duration for single stimulus video quality assessment is 5 seconds [21] - the shortest duration that provides satisfactory levels of accuracy. Consequently all 240 test sequences were cut (if necessary) to 5 seconds.

C. Experimental Setup

A Panasonic BT-4LH310 LCD reference monitor with a peak luminance of 210 cd/m^2 (measured using a Konica Minolta CS-2000 spectroradiometer), a contrast ratio of 400:1, 3840×2160 spatial resolution (measuring 65.4×36.8 cm), BT.2020 colour space [1] (full range), and a refresh rate of 60 fps was used. The display was connected via quad 3G-SDI to a Windows PC running Matlab 2017b. The viewing distance was chosen as 1.5 H - the optimal for the tested resolution [22]. The viewing environment conformed to the home environment conditions outlined in BT.500-13 [15].

D. Testing Methodology

The experiment was conducted in two phases to reduce the length of the experiment: the first phase contained the sequences captured by ourselves; the second contained the 12 sequences from the Harmonic and Netflix video

databases.

Prior to the experiment, each participant took part in a brief training session to acclimatise themselves with the testing process. A complete session lasted no longer than 30 minutes, and involved viewing the 120 test sequences for that phase using a single stimulus methodology. Each trial consisted of the participant viewing a 3 second mid-level grey screen before viewing a randomly selected sequence. Participants' then recorded their opinion on a continuous quality scale from 0 to 5 [15]. A single-stimulus, rather than a double-stimulus methodology was chosen as it more similar to typical viewing environments i.e. comparisons with a high quality reference sequence are generally not possible.

E. Participants

Twenty-two participants from the University of Bristol were paid to take part in each phase (both expert and non-expert viewers). The average age ($\pm\sigma$) of participants for the first and second phase was 30$\pm$8 (15 male, 7 female) and 29$\pm$7 (13 male, 9 female) respectively. Twelve participants took part in both phases. All participants had normal or corrected-to-normal colour vision (verified with a Snellen chart).

IV. Results and Discussion

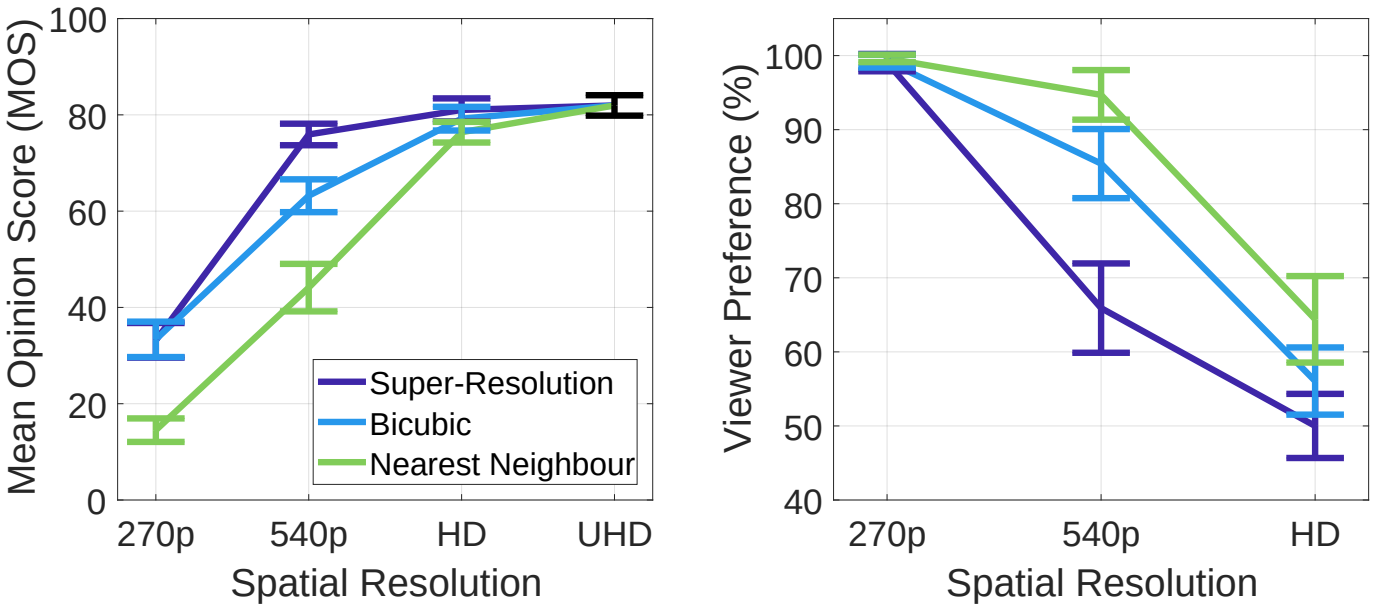

Fig. 3: Results from the subjective experiment showing the relationship between visual quality (left) and viewer preference (right) with spatial resolution. Error bars represent the 95% confidence interval of the mean.

A. Overall Performance

Mean Opinion Scores (MOS) were calculated for each test condition and linearly scaled to the range 0-100 (bad to excellent). The results from the experiment can be viewed in Fig. 3 (left), in which the average MOS of all test sequences is reported. The results show that visual quality (MOS) increases with spatial resolution. A one-way repeated measure ANOVA with Greenhouse-Geisser correction (due to violation of the sphericity assumption) confirms that resolution has a significant effect ($p<0.05$) on MOS for the three filters:

- Super-Resolution: $F(1.3, 30.5) = 400$, $p \approx 0$
- Bicubic: $F(1.4, 33.0) = 325$, $p \approx 0$
- Nearest Neighbour: $F(1.3, 29.4) = 496$, $p \approx 0$

A one-tailed paired t-test at a 0.025 significance level can be used to assert whether there are significant differences in MOS between the adaptation filters. The results in Table II) show that the super-resolution filter leads to significant increases in visual quality over nearest neighbour and bicubic.

Even though increased spatial resolutions have the ability to resolve higher spatial frequencies, and therefore reduce the amount of visible aliasing energy. The results indicate that any subsequent perceptual benefits diminish -

TABLE I: The results when using a one-tailed paired t-test to compare MOS between the adaptation filters. A '1' indicates that the filter in that row is statistically superior to the filter in the column (the opposite holds for '-1').

Filter	Nearest Neighbour	Bicubic	VDSR
Nearest Neighbour	-	-1	-1
Bicubic	1	-	-1
VDSR	1	1	-

especially beyond HD. There are two predominant reasons for this: natural images generally follow a $1/f$ spectral distribution [23], and spatial frequencies are attenuated by the human visual system (the higher the frequency, the greater the attenuation) [24]. This culminates in visible aliasing energy exponentially decreasing as spatial resolution increases, up to the acuity limits of the visual system (around 32 cycles per degree) [25]. Beyond this point optical reality is simulated, and it is therefore futile to increase spatial resolution any further without larger displays and/or reduced viewing distances.

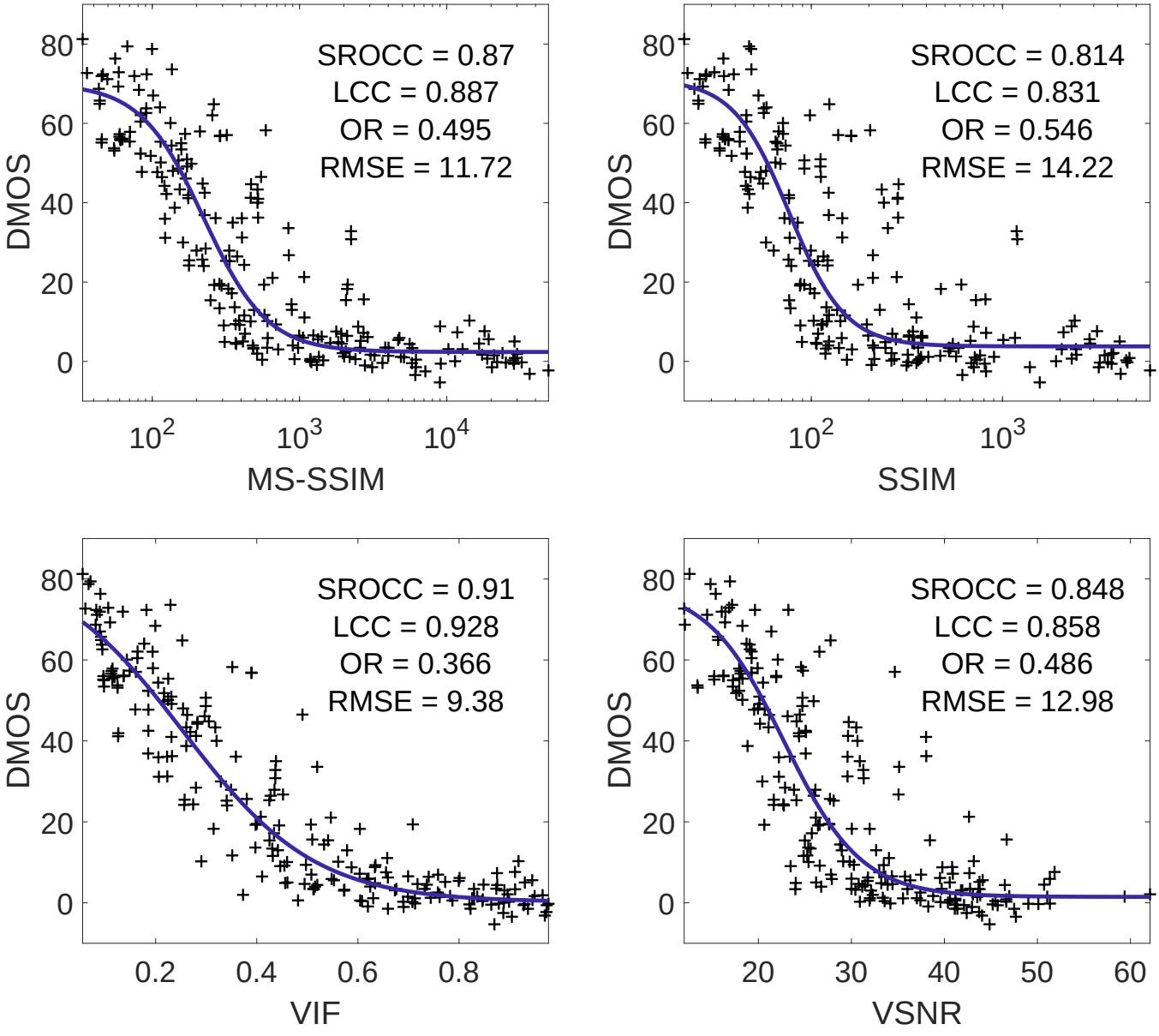

Fig. 4: The relationship between the image quality metric predictions and DMOS. The (blue) line is the four parameter logistic fitting curve.

B. Viewer Preference

By calculating the percentage of participants who gave UHD-1 resolution a higher subjective rating than the lower test resolutions for each sequence, we can estimate viewer preference for UHD-1 content. Fig. 3 (right) shows that the average viewer preference for UHD-1 increases as spatial resolution decreases, and that the type of filter has a large impact.

Whereas almost every participant preferred UHD-1 to 270p, only 50% of participants on average preferred UHD-1 to HD when using the super-resolution filter - indicating negligible perceptual differences between the two resolutions. In order to confirm this assertion, a one-tailed Wilcoxon signed rank test[2] was used to compare the raw opinion scores of the UHD-1 and HD resolutions for all three filters:

- Super-Resolution: $Z = 1.37$, $p = 0.085$
- Bicubic: $Z = 4.36$, $p \approx 0$
- Nearest Neighbour: $Z = 8.31$, $p \approx 0$

These results demonstrate that while there is a significant ($p<0.025$) reduction in visual quality between UHD-1 and HD resolutions for the bicubic and nearest neighbour filter, there is no significant reduction for the super-resolution filter.

This suggests that UHD-1 content could be down-sampled to HD resolution for storage/transmission, and then up-sampled at the receiver using the super-resolution filter without significantly affecting visual quality. While the introduction of video compression would likely affect these findings [19], the super-resolution method [10] could be trained on compressed content in an attempt to alleviate this issue.

C. Quality Metrics

Future adaptive formats, in which optimal resolutions are selected given a set of constraints (channel conditions, desired quality etc.), will require accurate quality metrics that are robust to content dependence. We will therefore investigate whether current image quality metrics can successfully model visual quality across a range of test conditions.

The image quality metrics considered were: MS-SSIM [26], PSNR [27], SSIM [28], VIF [29] and VSNR [30]. It is assumed that the distortions which arise during spatial resolution adaptation will be temporally consistent, and therefore any marginal gains associated with video quality metrics will be offset by the increase in computational complexity.

The predictions of the quality metrics were averaged over all 300 frames, and then subsequently fitted with a logistic function to reduce non-linearities [31]. Spearman Rank Correlation (SROCC), Pearson Linear Correlation (LCC), Outlier ratio (OR) and Root Mean Squared Error (RMSE) [31] were used to appraise accuracy, monotonicity and consistency of the predictions. Differential mean opinion scores (DMOS) [15] were calculated by subtracting the MOS of the lower resolution from the MOS of the UHD-1 reference.

TABLE II: F-test results for the quality metrics at a 95% confidence interval. A '1' indicates that the metric in that row is statistically superior to the metric in the column (the opposite holds for '-1'), while a '0' indicates that there is no statistically significant difference between the two metrics.

Metric	PSNR	VSNR	SSIM	MS-SSIM	VIF
PSNR	-	0	0	-1	-1
SSIM	0	-	0	-1	-1
VSNR	0	0	-	0	-1
MS-SSIM	1	1	0	-	-1
VIF	1	1	1	1	-

The SROCC, LCC, OR and RMSE values for PSNR are: 0.820, 0.841 0.537 and 13.68 respectively. The performance of the other quality metrics is shown in Fig. 4, alongside the relationship between the predictions and DMOS.

All tested quality metrics report high correlation coefficients (>0.8) and compact predictions around the fitting curves (OR). This suggests that they can successfully characterise visual quality over a range of spatial resolutions and adaptation filters. Table II reports F-test [32] results between the quality metrics, and demonstrates that the

[2]The normality assumption of a paired t-test was violated.

predictions of VIF are statistically superior to all other tested quality metrics - thus indicating that it would be the most suitable to select spatial resolutions within future adaptive formats.

V. Conclusions

This paper presents a publicly available video database (BVI-SR) which contains video sequences with a range of spatial resolutions up to UHD-1. This database is used for a large-scale subjective experiment that characterises the relationship between spatial resolution and visual quality. Results show that while spatial resolution has a significant impact on visual quality, there is no significant difference in subjective scores between UHD-1 and HD resolutions when using a state-of-the-art CNN-based super-resolution method. We further demonstrate that image quality metrics can successfully model visual quality across the range of test conditions, and therefore could be utilised within future adaptive formats.

REFERENCES

[1] ITU-R Recommendation BT.2020-2, "Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange," 2015.

[2] D. Bull, *Communicating Pictures: A Course in Image and Video Coding*, Elsevier, 2014.

[3] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P. Le Callet, "Comparing upscaling algorithms from HD to Ultra HD by evaluating preference of experience," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 208–213.

[4] P. Hanhart, P. Korshunov, T. Ebrahimi, Y. Thomas, and H. Hoffmann, "Subjective quality evaluation of high dynamic range video and display for future TV," 2014.

[5] A. Mackin, F. Zhang, and D. Bull, "A study of subjective video quality at various frame rates," in *Image Processing (ICIP), 2015 22nd IEEE International Conference on*, 2015.

[6] A. Mackin, K. Noland, and D. Bull, "The visibility of motion artifacts and their effect on motion quality," in *Image Processing (ICIP), 2016 23rd IEEE International Conference on*, 2016.

[7] G. Van Wallendael, P. Coppens, T. Paridaens, N. Van Kets, W. Van den Broeck, and P. Lambert, "Perceptual quality of 4K-resolution video content compared to HD," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6.

[8] A. Mackin, F. Zhang, M. Papadopoulos, and D. Bull, "Investigating the impact of high frame rates on video compression," in *Image Processing (ICIP), 2017 24th IEEE International Conference on*, 2017.

[9] A. Mackin, K. Noland, and D. Bull, "High frame rates and the visibility of motion artifacts," *SMPTE Motion Imaging Journal*, vol. 126, no. 5, pp. 41–51, 2017.

[10] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[11] Harmonic, "4K Demo Footage," www.harmonicinc.com/4k-demo-footage-download/.

[12] Netflix, "Video test media," https://media.xiph.org/video/derf/.

[13] ITU-R WP6Q, "SAMVIQ - Subjective assessment methodology for video quality," September 2003.

[14] J. Lee, L. Goldmann, and T. Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia tools and applications*, vol. 67, no. 1, pp. 31–48, 2013.

[15] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[16] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[17] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[18] K. Berger, Y. Koudota, M. Barkowsky, and P. Le Callet, "Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[19] M. Afonso, F. Zhang, A. Katsenou, D. Agrafiotis, and D. Bull, "Low complexity video coding based on spatial resolution adaptation," in *Image Processing (ICIP), 2017 24th IEEE International Conference on*, 2017.

[20] S. Winkler, "Analysis of public image and video databases for quality assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 616–625, October 2012.

[21] F. Moss, C. Yeh, F. Zhang, R. Baddeley, and D. Bull, "Support for reduced presentation durations in subjective video quality assessment," *Signal Processing: Image Communication*, vol. 48, pp. 38–49, 2016.

[22] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[23] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: computation in neural systems*, vol. 14, no. 3, pp. 391–412, 2003.

[24] P. Barten, *Contrast sensitivity of the human eye and its effects on image quality*, SPIE press, 1999.

[25] A. Watson, "High frame rates and human vision: a view through the window of visibility," *SMPTE Motion Imaging Journal*, vol. 122, no. 2, pp. 18–32, 2013.

[26] Z. Wang, E. Simoncelli, and A Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. IEEE, 2003, vol. 2, pp. 1398–1402.

[27] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.

[28] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[30] D. Chandler and S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE transactions on image processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[31] International Telecom Union, "*Tutorial: Objective perceptual assessment of video quality: Full reference television*," 2004.

[32] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.