Study of high dynamic range video quality assessment

Manish Narwaria, Matthieu Perreira da Silva, Patrick Le Callet

Study of High Dynamic Range Video Quality Assessment

Manish Narwaria, Matthieu Perreira Da Silva and Patrick Le Callet

LUNAM University - IRCCyN CNRS UMR 6597, 44306, Nantes, France

ABSTRACT

In recent years, High Dynamic Range (HDR) imaging has attracted significant attention from industry and academia. As a result, there are currently several on-going efforts towards standardization and benchmarking of existing tools for HDR image and video, and one of the key aspects is that of video quality measurement (both subjective and objective approaches). Therefore, this paper aims to identify few key challenges in the said area and then discuss existing solutions. Specifically, we first discuss a few important practical aspects that make HDR video quality measurement potentially challenging. Second, we report our recent efforts towards developing HDR video datasets that have been subjectively annotated for visual quality. Finally, we analyze and compare the effectiveness of existing solutions for objective quality prediction.

Keywords: HDR video quality, subjective and objective methods, temporal factors

1. INTRODUCTION

The explosion in the number of computers and digital systems connected by networks such as the Internet has brought a flow of instant information into a large and increasing number of homes and businesses. This has greatly facilitated the delivery of multimedia signals to the end-users, resulting in ever increasing levels of interaction with multimedia information. Such effect is not merely quantitative in nature but has had significant impact on the qualitative aspects. Thus, todays' users are not just constantly in touch with multimedia information through a wide range of devices but more importantly the nature of such engagement has continuously evolved. It means that the users do not simply consume multimedia information but also expect a lot more immersion and interaction with the content. While the study of user experience and behavior with regards to interaction with multimedia content is complex and challenging, it is generally agreed that visual signals (video) are important[1] to what constitutes and defines the end user experience. Hence, there have been increased efforts (from view point of research as well technology deployment) to develop and standardize video technologies which can provide higher levels of engagement or immersion to the end-user, in comparison to the traditional video technologies. Some of these include ultra-high definition (UHD), High Frame Rate (HFR), 3D TV, and High Dynamic Range (HDR) imaging. The first two technologies primarily aim at increasing spatial and temporal resolution in order to capture extra details and improve temporal coherency. On the other hand, while 3D technologies rely on the additional dimension of depth in order to bring realism, HDR imaging attempts to improve video contrast and level of details in order to enhance immersion. It has, in particular, attracted significant research attention in recent times because it depends on a more realistic representation of the scene in terms of physical quantities, and can thus potentially bring more realism into video viewing experience of the end-user. Stated differently, HDR imaging technologies aim to overcome the inadequacies of the traditional low dynamic range (LDR) capture and display technologies via better video signal capture, representation and display, so that the dynamic range of the video can better match the instantaneous range of the eye. Despite the limitations that exist especially with regards to the display of native HDR videos, HDR imaging is an active research area.

Thus, there are current efforts in order to enable HDR imaging at professional and consumer levels. However, as is the case with every technology, the operational, technical and management issues throughout the video delivery chain have to effectively addressed in the context of HDR imaging. One of the most prominent of these is that of controlling and calibrating perceptual video quality, at different points along the video delivery chain. This is more so in the light of the fact that there has been a paradigm shift from the traditional Quality of Service (QoS, where the issues are generally prioritized from the service provider view point) to Quality of Experience (QoE, where the expectations of the end user are more important) centered HDR content delivery. Subjective quality estimation of HDR content has been addressed by earlier works,[2–6] although many of these used image stimuli. Likewise, few works[7–10] attempted to address the problem of objective evaluation of HDR

quality though most focus on still images (except HDR-VQM[9] which has been designed for video). As opposed to all these works, the main aim of this paper is to highlight practical issues that can be encountered in both subjective and objective assessment of HDR video quality quality. This will provide insights into the specific issues that one should consider while dealing with the problem. In addition, we also provide experimental results and compare the performance of a few objective methods in order to understand their strengths and limitations for HDR video quality prediction.

2. HDR VIDEO QUALITY: ISSUES AND CHALLENGES

Video quality estimation, in general, is necessary in video processing systems in order to optimize, calibrate and benchmark algorithms. Video quality can be assessed by both subjective and objective methods. The former involves the use of human subjects to judge and rate the quality of the test stimuli. With appropriate laboratory conditions and a sufficiently large subject panel, it remains the most accurate method. The latter quality assessment method employs a computational (mathematical) model to provide estimates of the subjective video quality. While such objective models may not mimic subjective opinions accurately in a general scenario, they can be reasonably effective in specific conditions/applications.

Pertaining to HDR video quality, it may be pointed out that its analysis and measurement involves a few added issues, in comparison to traditional video quality estimation. This is true for both subjective and objective approaches, and the following sections elaborate on them.

2.1 Factors in subjective estimation of HDR video quality

In this section, we discuss a few important issues that are encountered in subjective HDR video quality assessment and these should be carefully considered in order to obtain reliable, and reproducible subjective measurements.

- **Video representation:** It is useful to highlight that HDR video data is typically only proportional to physical luminance values which characterized the scene and not equal to it. More specifically, unless there is a prior and accurate camera calibration, luminance values in an HDR video file represent the real world luminance up to an unknown scale*. As a consequence, care should be taken while using HDR data directly for calibration purposes.

- **Source content selection:** Similar to the case of traditional subjective video quality estimation, selecting a set of source (reference) video stimuli which, under a given context and/or application scenario, can challenge specific aspects of the algorithm under investigation, is necessary. For eg., in case of evaluating and validating a video coding method, it is necessary that the source content is selected such that it challenges the codec in terms of its ability to cope with both spatial and temporal redundancy. However, in case of HDR, the selection of source video content can involve an additional dimension related to the dynamic range of the content. For example, scenes with different dynamic range must be selected to challenge the algorithm under consideration. The classical definition of dynamic range (ratio of maximum and minimum luminance) can be used but its important to remember that it may suffer from drawbacks such as susceptibility to outliers and can be misleading in some situations (eg. a tiny patch of very dark and bright pixels can inflate the dynamic range). More sophisticated and recently proposed solutions (eg. the one in Ref. 11) could also be used as alternative index for source content selection based on certain perceptual considerations. Thus, HDR source video selection should be viewed as a multi-dimensional problem where in the traditional measures such as spatial and temporal information[12] should be complemented with additional HDR specific information.

- **Requirement of specialized display:** Ideally, the visualization of HDR video will require displays whose contrast ratio and peak luminance can match that of the real scene. This is, of course, practically not feasible (since the contrast and luminance found in real world can easily exceed 10 orders of magnitude which is

*Even with calibration, the HDR values represent real physical luminance with certain error. This is because the camera spectral sensitivity functions which relate scene radiance with captured RGB triplets cannot match the luminous efficiency function of the human visual system.

not possible to be achieved with current display technologies) nor desirable (because the instantaneous range of human vision is about 5 orders of magnitude[†]). Thus, displays that can cover the instantaneous vision range are typically used. The peak brightness in such displays will depend on the content but can generally display a maximum brightness up to 4000-5000 cd/m^2.

- **Video rendering:** Accurate rendering of HDR video on an HDR display is non-trivial due to two specific reasons. First, since HDR values are related to the actual luminance, the maximum luminance is scene (content) specific. Hence, there is no fixed white point and the HDR values must be interpreted based on the display used to view the HDR video. Second, since the maximum HDR display luminance and contrast are usually lower than the actual scene, even HDR display require range-reduced (or tone mapped) HDR video signal. This introduces another variable in the process of rendering which can modify the factors such as the visibility of details especially in very dark and very bright areas, artifact visibility and temporal coherency, thereby affecting the overall appearance of the video. In other words, the HDR video must be graded according to the target display in order to preserve the artistic intent captured in the HDR content. Hence, subjective evaluation of HDR video quality will almost always be dependent both on the display characteristics and the rendering method adopted, and this should be kept in mind for sake of reproducibility of subjective results as well as evaluation of artifacts due to a specific algorithm under test. In contrast to this, such considerations are minimal in case of the traditional LDR video quality measurement.

- **Viewing conditions:** HDR video viewing will involve higher levels of contrast and brightness. Thus, proper ambient lighting settings are necessary to minimize visual discomfort. Improper settings can result in maladapted viewing conditions which can hamper visibility of details. This is especially critical for videos since there will be a continuous variation of luminance levels. The exact illumination values should ideally depend on the maximum luminance considered in HDR video rendering, and may also depend on the content. Ambient light setting should in general be decided based on viewer comfort and adaptation levels.

- **Paired comparison tests:** Care must be taken that the HDR stimuli to be compared are at same (or at least similar) luminance levels,and the ambient light is set accordingly. Thus, comparing two stimuli from different source content via paired comparison can be tricky especially for HDR videos where the luminance could vary greatly over time. Studying the impact of tone mapping is an interesting use case where observers watch both HDR and LDR stimuli simultaneously. Since the peak brightness of the displays can be very different, arriving at a comfortable illumination level is not easy. An alternative is to use higher illumination around the HDR display while the diffused light can act as the illumination source for the LDR display.

2.2 Objective assessment of HDR video quality

The main challenges in objective estimation of HDR video quality stem from the representation of visual information in HDR videos. Recall that unlike the traditional videos which store gamma corrected pixels, HDR pixels are related to physical luminance. Hence, any objective approach should consider this aspect. In literature, there are two main strategies that have been adopted to tackle this.

The first one is HVS-like (human visual system) approach. This is, therefore, an *ideal* approach in that the luminance is processed similar to what the human eye does. Indeed, HVS based quality assessment methods such as the Visual Difference Predictor (VDP)[13] and the more recent HDR-VDP-2[8] predict quality based on luminance. In case of the former, LDR values are first converted to physical luminance by using a display model. Of course, for obvious reasons complete HVS modeling is neither possible nor practical. However, some of the well-studied concepts of the HVS that play a role in quality judgment can still be approximated by computational models. Thus, methods such as VDP and HDR-VDP-2 attempt to mimic some functionalities of the HVS that

[†]Given sufficient adaptation time, the dynamic range of human eyes is about 13 orders of magnitude. However, since the typical frequency in video signals does not allow sufficient adaptation time, the dynamic vision range (5 orders of magnitude) is more relevant in the context of this paper as well as HDR video processing in general.

are relevant for quality judgment. These can include modeling of the intra-ocular light scatter, photo receptor spectral sensitivity curves, frequency and orientation selectivity, luminance masking and contrast sensitivity, to list a few. The said modeling employs mathematically tractable functions (for instance in HDR-VDP-2 a modulation transfer function is used to model light scattering) that can approximate a particular behavior of the HVS to luminance.

The second approach is based on first converting physical luminance into values that are closer to the perceived luminance, and then using an LDR method to compute objective quality. Thus, unlike the first approach, the goal is to derive a mapping function that can transform physical luminance values to another space where the differences in values can be perceived linearly. For instance, a simple logarithmic transformation can be useful since it saturates at high luminance. This means that changes in high luminance will be less noticeable, and this is something which is approximately in line with HVS's response to luminance. Another luminance transformation function proposed by Aydin et al.[14] is based on the idea that the transformed values should be close to the dynamic range of a typical cathode ray tube display in mean squared error sense. This in turn means that the transformed luminance values can be treated as *similar* to the ones that are gamma encoded i.e. perceptually uniform, and hence can be more effective towards for computing objective quality. As can be noted, this approach allows the subsequent use of any LDR quality assessment method whose input will be transformed luminance values.

While the two approaches just discussed have their own advantages and disadvantages, it is more important to remember that the input to the objective method should be calibrated according to the display on which the HDR video is meant to be viewed. This might involve simple processing such as frame-by-frame linear scaling, temporal linear scaling (eg. this was employed in Ref. 9) or more sophisticated one based on dual-modulation. By contrast, objective assessment of LDR video quality does not generally require such display specific processing.

3. HDR VIDEO DATASET

This section describes the details of the subjective study that we carried out in order to obtain ground truth for further analysis of HDR video quality. Our aim was to create a dataset which is diverse, both in terms of content and processing.

3.1 Test Material Preparation

We used 9 source (reference) HDR sequences, and the first frames from each of them are shown in Figure 1. Three of these sequences (src1, src2 and src3) were made available as part of our project[‡]. Five sequences (src4 to src8) were obtained from Ref. 15, and one sequence (src9) was computer-generated using dedicated software[§]. The frame resolution was full HD (1920 by 1080 pixels) and the frame rate was 25 frames per second (fps). The dynamic range for each source sequence is also indicated in Figure 1. It was computed as the maximum of the dynamic ranges of individual frames i.e.

$$Dynamic\,Range = max\left\{ log_{10}\Big(\frac{L_{max}^{(t)}}{L_{min}^{(t)}}\Big)\right\}_{t=1,2\ldots F} \tag{1}$$

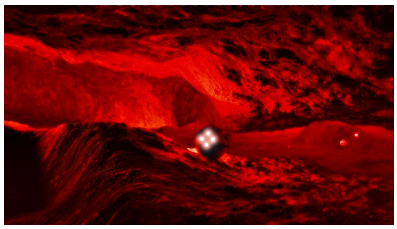
where $L_{max}^{(t)}$ and $L_{min}^{(t)}$ respectively denote maximum and minimum relative luminance values for frame t, and F being the total number of frames. For robustness against outliers and avoiding division by zero, these were computed after discarding 0.1% brightest and darkest values. The resulting dynamic range for each source sequence is also indicated in Figure 1. Although we report the video dynamic range as the maximum dynamic range of a frame in the sequence, we found that all the frames had dynamic range in excess of 3 for all the source sequences, except src5 and src6 in which the percentages of such frames were 90.54% and 53.51% respectively. Since most LDR displays cannot display more than 3 orders of magnitude, the video content used by us requires an HDR display.

[‡]NEVEx project FUI11, related to HDR video chain study.

[§]This sequence has been generated within the framework of the project UHD4U, related to the study of immersive video technologies including ultra-high definition and HDR.

(a) src1, *Dynamic Range* = 6.87

(b) src2, *Dynamic Range* = 6.53



(c) src3, *Dynamic Range* = 6.26

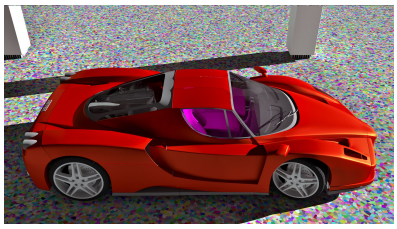
(d) src4, *Dynamic Range* = 4.42

(e) src5, *Dynamic Range* = 4.45

(f) src6, *Dynamic Range* = 3.98

(g) src7, *Dynamic Range* = 4.11

(h) src8, *Dynamic Range* = 4.15



(i) src9, *Dynamic Range* = 7.44

Figure 1: Tone mapped versions of first frame of the 9 reference (source) HDR video sequences (src1 to src9) used in the subjective study. The number of frames in src1 to src9 were respectively 200, 200, 200, 405, 317, 271, 402, 371 and 250.

The source sequences were compressed using a backward-compatible (in our context backward compatibility with existing standard 8-bit displays) HDR video compression method. We do not discuss the method here, but for the purposes of this paper it is sufficient point out that in general, any backward-compatible HDR compression scheme comprises of 3 main steps: (a) forward tone mapping (TMO) in order to convert HDR video to LDR (8-bit precision), (b) compression and decompression of the LDR video by a standard LDR video compression method, (c) inverse tone mapping (iTMO) of the decoded (decompressed) LDR bit stream to reconstruct HDR video (additional enhancement such as those based on residuals can also be employed). In this paper, the LDR video was encoded and decoded using HEVC at different bit rates. For the subjective viewing tests, we selected 8 bit rates (by varying the quantization parameter QP) such that the resultant HDR video quality covered the entire rating scale, i.e., from excellent (rating 5) to bad (rating 1). With the inclusion of the source sequences, we obtained a total of 153 HDR video sequences (9 src $\times$ 8 bit rates $\times$ 2 TMO-iTMO processing + 9) to be rated by the subjects. Regarding TMO-iTMO processing, one was spatial only (i.e. TMO-iTMO applied on each frame independently) while the other applied a temporal correction factor as proposed in Ref. 16. Thus, for each source sequence, there were 8 distorted sequences from each of the TMO-iTMO processing.

3.2 Rating methodology

Our study involved 25 paid observers who were not expert in image or video processing. They were seated in a standardized room conforming to the International Telecommunication Union Recommendation (ITU-R) BT500-11 recommendations.[17] Prior to the test, observers were screened for visual acuity by using a Monoyer optometric table and for normal color vision by using Ishiharas tables. All of them had normal or corrected to normal visual acuity and normal color perception. For rating the test stimuli, we adopted the absolute category

Table 1: C_p values per source (src) sequence. There are 16 distorted sequences from each source, 8 from each TMO-iMO processing (spatial and temporal).

	MRSE	P-PSNR	P-SSIM	HDR-VDP-2.2	P-VIF	HDR-VQM
src1	0.9353	0.9630	0.9403	0.9572	0.9770	0.9251
src2	0.8954	0.9156	0.9865	0.9939	0.9762	0.9935
src3	0.9234	0.9773	0.9739	0.9920	0.9742	0.9931
src4	0.9759	0.9899	0.9899	0.9833	0.9366	0.9885
src5	0.9763	0.9589	0.9472	0.9794	0.9375	0.9844
src6	0.9426	0.9466	0.9810	0.9328	0.9479	0.9880
src7	0.9857	0.9898	0.9624	0.9838	0.9842	0.9922
src8	0.9705	0.6790	0.9864	0.9789	0.8979	0.9328
src9	0.6211	0.4190	0.8625	0.9929	0.8294	0.9118

rating with hidden reference (ACR-HR), which is one of the rating methods recommended by the ITU in Rec. ITU-T P.910.[12] The ACR-HR is a category judgment method where the test stimuli are presented one at a time and rated independently on a category scale. The rating method also includes the source sequences (i.e. undistorted) to be shown as any other test stimulus without informing the observers. This is therefore termed a hidden reference condition, and the advantage is that it implicitly encourages the observers to rate video quality and not fidelity since there is no reference to compare with. To quantify the video quality, a five-level scale is used: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor) and 1 (Bad). We chose a discrete five-level scale because it is more suitable for naive (non-experts in image processing) observers, and it is easier for them to quantify the quality based on an adjective ('Excellent', 'Good', 'Fair', 'Poor' and 'Bad'). We also employed post-experiment screening of the subjects in order to reject any outliers in accordance with the Video Quality Experts Group (VQEG) multimedia test plan, and in our case one observer was rejected.

3.3 Display

For displaying the HDR video sequences, SIM2 Solar47 HDR display was used which has a maximum displayable luminance of 4000 cd/m^2. The ambient light composed of light emitted by a source (at 100 cd/m^2) placed above the display. The viewing distance was set to three times the height of the screen (active part), i.e., approximately 178 cm. For rendering the HDR videos, we employed a dual modulation algorithm by taking into account the display point spread function (PSF), color correction, and also preserving the temporal coherency in the rendered video.

4. EXPERIMENTAL RESULTS AND ANALYSIS

We now compare the performance of few objective methods including HDR-VQM,[9] VIF,[18] PSNR and SSIM. All these methods operate on the perceived luminance values. In addition, we considered two other methods which predict quality based on display-referred luminance values. The first one is the HDR-VDP-2 originally proposed in Ref. 8, and we employed its re-calibrated version HDR-VDP-2.2.[10] The second method is the MRSE (Mean Relative Squared Error) which is a variant of the traditional mean squared error, MSE, in that it normalizes the error by the magnitude of luminance at each point in order to account for reduced sensitivity at high luminance. Since all the considered objective methods are full reference, the performance was evaluated using only the 144 distorted sequences (i.e. excluding the source sequences).

4.1 Correlation analysis

As the dataset used in this paper focuses on HDR video compression, we report the performance of different methods based on each source sequence, in order to study how they behave according to content. Recall that in our dataset, we employed 8 different bit rates and 2 different TMO-iTMO processing (first one is spatial only while the second one is temporal). As a result, we have 16 distorted sequences (8 each from the two TMO-iTMO processing) for each reference sequence. We first present, in Table 1 the correlation values of different objective methods with the subjective scores per source sequence. We can see that while all the methods appear to provide relatively high correlation, the performance is content-dependent especially for methods like PSNR and MRSE.

On the other hand, methods like HDR-VQM and HDR-VDP-2.2 are relatively more consistent across different source content. It is also interesting to point out that the correlation values for few objective methods (P-PSNR, MRSE and P-VIF) were found to be higher and more consistent when considering each TMO-iTMO processing separately. This suggests that the performance of such methods can suffer even for the same source content (and codec which was HEVC in this case) as the number of processing increase. Such observation is perhaps more relevant in the context of HDR video processing where other processing such as TMO, iTMO (inverse TMO), display processing etc. can affect video quality, in addition to the distortions induced by the compression algorithm.

4.2 Analysis based on prediction errors and outliers

The second part of the analysis is based on the number of outliers and the prediction error measured in terms of the root mean squared error. Outlier analysis is another approach to evaluate objective methods for their prediction accuracy. It is different from the usual correlation based comparisons in that it does not penalize the error that is within the limit of uncertainty in the subjective rating process. The main advantage of outlier analysis is that it helps to evaluate metric accuracy by taking into account the variability or uncertainty in subjective opinions, which are ignored in correlation based comparisons. Particularly, it can be very useful in applications such as video compression where one is generally interested in the rate distortion (RD) behavior of objective methods i.e. how the objective visual quality varies with bit rates for different source sequences and to what extent that compares with the subjective video quality.

In order to compute outliers and prediction errors, it is required that the objective scores be mapped (transformed) to the subjective scores via a mathematical function.[19] In our analysis, recall that we have 16 objective scores (from each objective method) and the corresponding subjective scores, for each source sequence. We also note from Table 1 that the objective scores indicate reasonably high linear relation with the subjective scores. Thus, we employed a two-parameter linear mapping function defined below, for the said transformation to avoid an over-fitted model (which can lead to near zero prediction errors for all the objective methods):

$$MOS_{objective} = a.\, objective_{score} + b \tag{2}$$

where $MOS_{objective}$ denotes the linearly transformed $objective_{score}$, and a, b are the parameters of such mapping which were computed by minimizing the squared error between $MOS_{objective}$ and corresponding MOS (subjective scores). Thus, we obtained $MOS_{objective}$ for each objective method and each source sequence.

To compute the number of outliers, we used the criterion specified in ITU-T P.1401.[19] It defines an objective prediction to be an outlier if the absolute prediction error between $MOS_{objective}$ and MOS is greater than the associated confidence interval $CI = \frac{z \times std(MOS)}{\sqrt{N_{observers}}}$, where $std(MOS)$ is the standard deviation of raw (i.e. per observer) scores, $N_{observers}$ denotes the number of observers and z depends on the considered distribution (in our case we assumed Student's t-distribution) and confidence level (95%). Likewise, we followed the definition in ITU-T P.1401[19] for computing the prediction error that takes into account the uncertainty in the subjective rating process, and penalizes errors that are bigger than the associated uncertainty (epsilon-insensitive RMSE). It is denoted as $RMSE^*$ in order to distinguish it from the traditional RMSE (which penalizes every error in prediction), and computed as

$$RMSE^* = \sqrt{\frac{1}{N-d}\sum_{n=1}^{N}(max(0, |MOS(n) - MOS_{objective}(n)| - CI(n))^2} \tag{3}$$

where N is the number of samples and d being the degrees of freedom in the mapping function. Thus, our case $N = 16$ (since we carried out the analysis per source sequence) and $d = 3$ (it is equal to one more than the number of free parameters in the mapping function).

Even though we computed the mapped objective values $MOS_{objective}$ for each source sequence considering 16 test conditions (8 bit rates$\times$ 2 TMO-iTMO processing), we plotted these values separately for each TMO-iTMO processing for sake of visual clarity. The plots for the first case i.e. spatial TMO-iTMO are presented in Figure 2 while the ones corresponding to the second case i.e. temporal TMO-iTMO are presented in Figure 3. Both

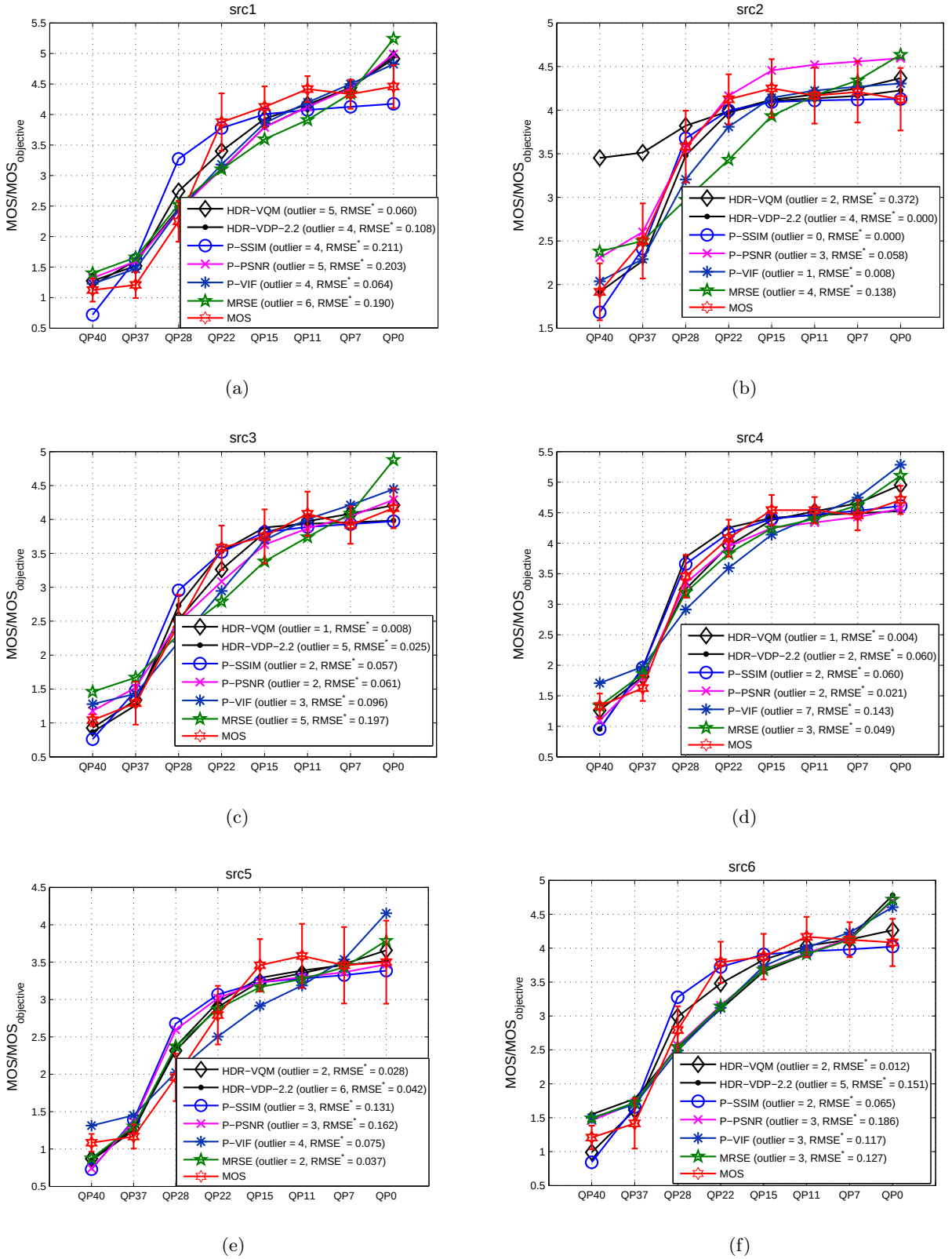
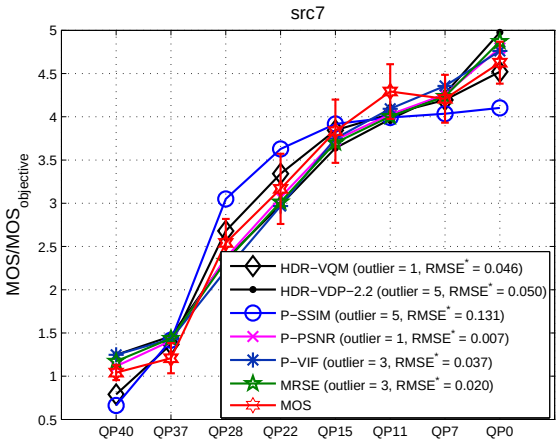
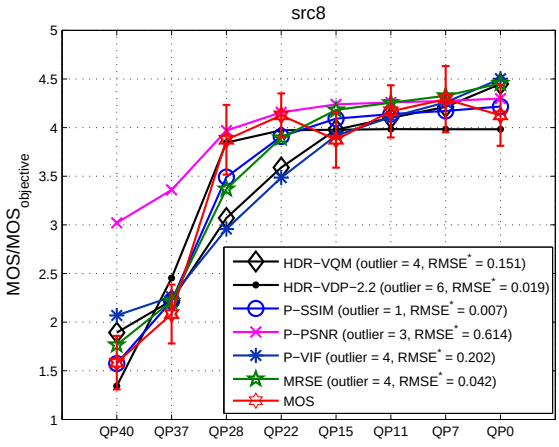


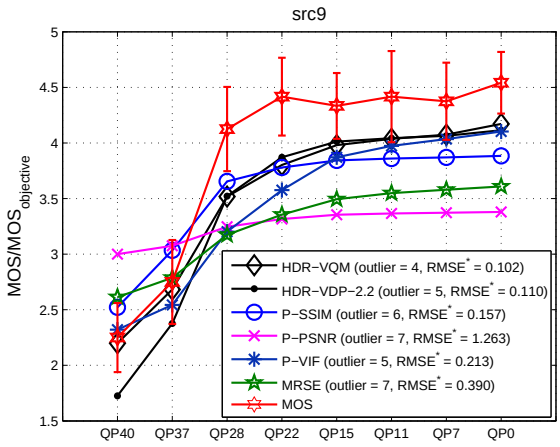
Figure 2: MOS (subjective scores) and $MOS_{objective}$ of different objective methods, for src1 to src6. The number of outliers and $RMSE^*$ values are also indicated in each plot. Error bars indicate the 95% confidence intervals for each corresponding MOS. These plots are for spatial TMO-iTMO.



(g)



(h)



(i)

Figure 2: (continued) MOS (subjective scores) and $MOS_{objective}$ of different objective methods, for src7 to src9. The number of outliers and $RMSE^*$ values are also indicated in each plot. Error bars indicate the 95% confidence intervals for each corresponding MOS. These plots are for spatial TMO-iTMO.

the figures provide a visual indication of how each objective method compares with the subjective scores, for the given source sequence and processing. Moreover, we have shown in these figures, the number of outliers as well as the $RMSE^*$ values for comparison. We can make the following observations from these plots, and the correlation values in Table 1 :

- The correlation values, number of outliers and $RMSE^*$ may not always be in agreement. For instance, the correlation values for P-VIF are nearly the same for src 1 to src3 but the corresponding number of outliers and $RMSE^*$ in Figure 2 are different in each case. Likewise, despite correlations being on a higher side, the number of outliers for some methods can exceed 50%. Hence, the results must be interpreted accordingly, and it would be more suitable to rely on all of them to make more informed judgments and conclusions about the prediction performance of objective methods.
- Another observation is that despite the same number of outliers, two or more objective methods can be distinguished by their $RMSE^*$ values. This is because $RMSE^*$ indicates the distance between $MOS_{objective}$ and MOS which exceeds the uncertainty, and obviously smaller will be better. For example, as indicated in Figure 3 for src7, the number of outliers for P-PSNR and MRSE is 1 but their $RMSE^*$ values are not equal (in this case, it is slightly smaller for P-PSNR). In contrast, the trend is reversed in case of src6 where both the methods result in 6 outliers but $RMSE^*$ value for MRSE is lower. Hence, the performance of these simpler methods is content-dependent.
- We also note that $RMSE^*$ values can be lower for an objective method despite a higher number of outliers. This happens because the method might give very accurate predictions for the points which are not outliers. For example, in Figure 3 for src4, HDR-VQM gives 3 outliers while HDR-VDP-2.2, P-SSIM and P-PSNR result in 2 outliers. Despite this, HDR-VQM has lower $RMSE^*$ than these methods.
- While we can note that the performance of different object methods can vary according to content, the results in Table 1 and Figures 2 and 3 indicate that overall HDR-VQM results in lower number of outliers as well as smaller $RMSE^*$ values.

5. CONCLUSIONS

HDR imaging has become popular within the video processing community primarily due to its high potential in offering the end-user a more immersive viewing experience. Hence, research and standardization efforts have been underway to deploy the technology at consumer levels. In the light of such activities as well as the foreseen technology push, management of HDR video quality will play an important role all along the video delivery chain. Thus, the aim of this paper was to highlight few crucial factors in subjective and objective estimation of HDR video quality, and highlight the differences with the traditional approaches. We also reported the performance of few objective methods on a set of HDR videos that were subjectively rated for their quality. Particular focus was placed on the analysis based on prediction accuracy for each source sequence. The results indicate that the performance f objective methods can be influenced by content, and this is true especially in case of methods like PSNR.

5.1 Acknowledgments

The authors wish to thank Romuald Pepion for his help in generating the subjective test results used in this paper. This work has been supported in part by the NEVEx project FUI11 financed by the French government, and the European project UHD4U.

REFERENCES

[1] Tang, J. and Isaacs, E., "Why do users like video?," *Computer Supported Cooperative Work (CSCW)* **1**(3), 163–196 (1992).

[2] Mantel, C., Ferchiu, S., and Forchhammer, S., "Comparing subjective and objective quality assessment of HDR images compressed with jpeg-xt," in [*16th IEEE International Workshop on Multimedia Signal Processing (MMSP)*], 1–6 (Sept 2014).

[3] Rerabek, M., Hanhart, P., Korsunov, P., and Ebrahimi, T., "Subjective and objective evaluation of HDR video compression," in [*Proc. Video Processing and Quality Metrics (VPQM)*], 1–6 (Jan. 2015).

[4] Narwaria, M., Perreira Da Silva, M., Le Callet, P., and Pepion, R., "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering* **52**(10), 102008–102008 (2013).

[5] Hanhart, P., Bernardo, M., Korshunov, P., Pereira, M., Pinheiro, A., and Ebrahimi, T., "HDR image compression: A new challenge for objective quality metrics," in [*Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*], 159–164 (Sept 2014).

[6] Narwaria, M., Perreira Da Silva, M., Le Callet, P., and Pepion, R., "Impact of tone mapping in high dynamic range image compression," in [*Proc. Video Processing and Quality Metrics (VPQM)*], 1–6 (Jan. 2014).

[7] Valenzise, G., Simone, F., Lauga, P., and Dufaux, F., "Performance evaluation of objective quality metrics for HDR image compression," in [*Proc. SPIE*], **9217**, 92170C–92170C–10 (2014).

[8] Mantiuk, R., Kim, K., Rempel, A., and Heidrich, W., "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," in [*ACM Trans. on Graphics*], **30**(4), 40, ACM (2011).

[9] Narwaria, M., Perreira Da Silva, M., and Le Callet, P., "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication* **35**(0), 46 – 60 (2015).

[10] Narwaria, M., Mantiuk, R., Perreira Da Silva, M., and Le Callet, P., "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging* **24**(1), 010501 (2015).

[11] Narwaria, M., Mantel, C., Da Silva, M., Le Callet, P., and Forchhammer, S., "An objective method for high dynamic range source content selection," in [*Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*], 13–18 (Sept 2014).

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

[13] Daly, S., "The visible differences predictor: an algorithm for the assessment of image fidelity," in [*Digital images and human vision*], 179–206, MIT Press (1993).

[14] Aydin, T., Mantiuk, R., and Seidel, H., "Extending quality metrics to full luminance range images," in [*Proc. SPIE*], **6806**, 68060B–68060B–10 (2008).

[15] Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., and Brendel, H., "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in [*Proc. SPIE*], **9023**, 90230X–90230X–10 (2014).

[16] Boitard, R., Bouatouch, K., Cozot, R., Thoreau, D., and Gruson, A., "Temporal coherency for video tone mapping," in [*Proc. SPIE*], **8499**, 84990D–84990D–10 (2012).

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

[18] Sheikh, H., Bovik, A., and de Veciana, G., "An information fidelity criterion for image quality assessment using natural scene statistics," *Image Processing, IEEE Transactions on* **14**, 2117–2128 (Dec 2005).

[19] "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," Recommendation ITU-T P.1401 (July 2012).

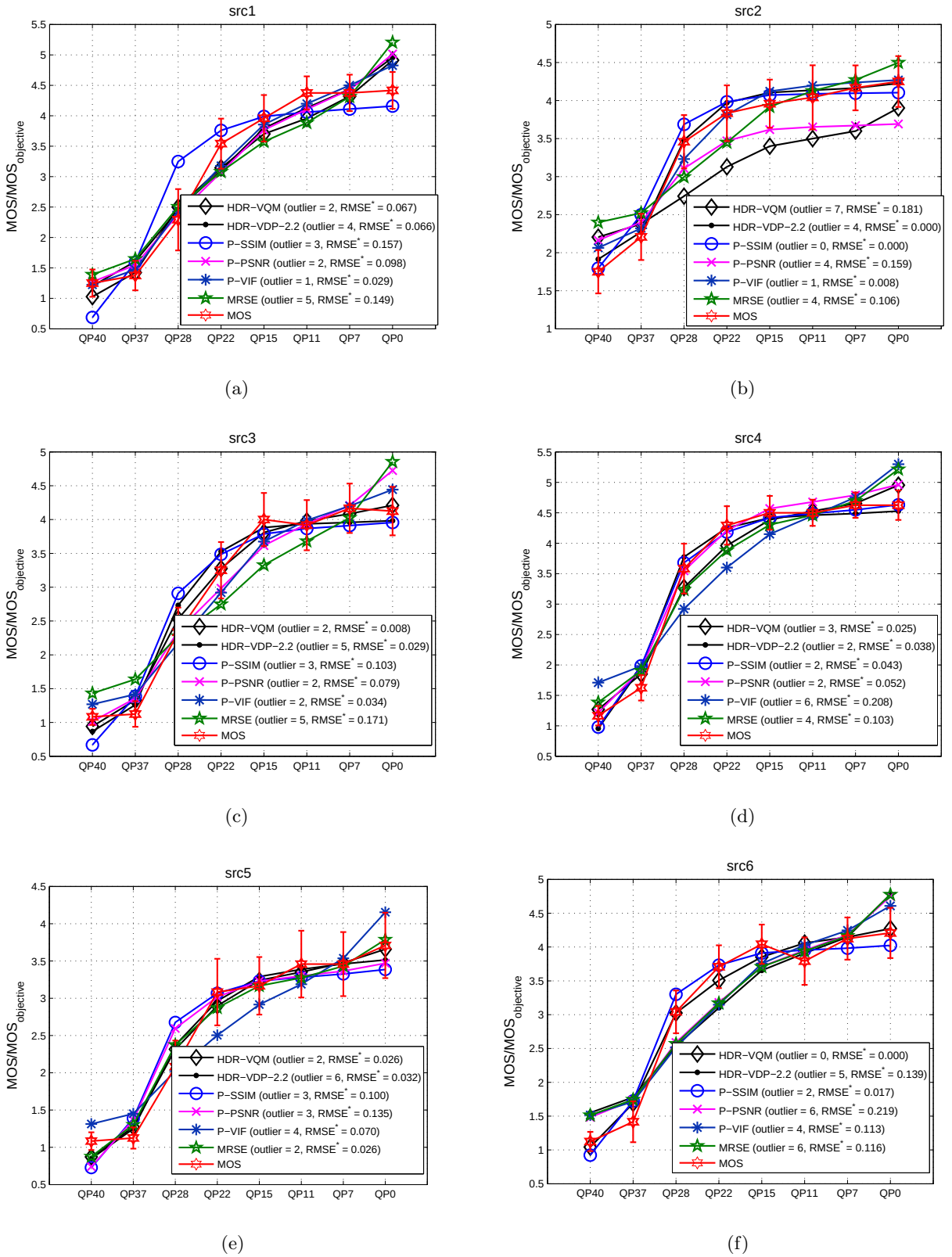
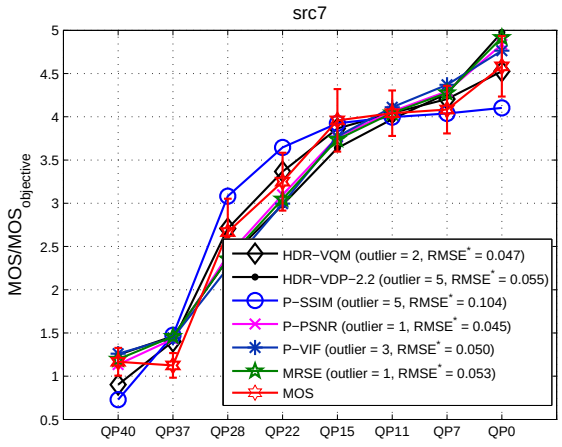
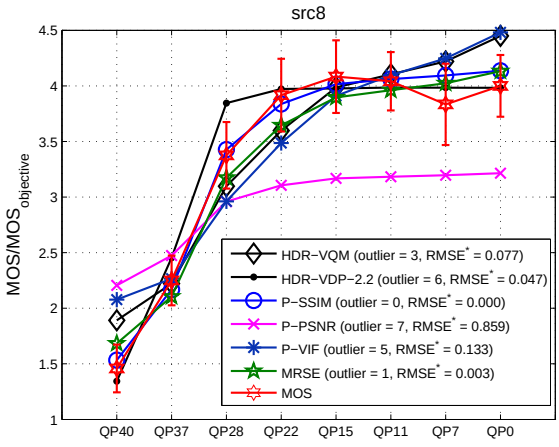


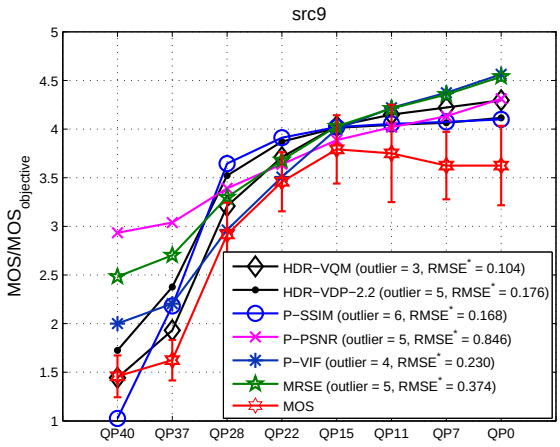
Figure 3: MOS (subjective scores) and $MOS_{objective}$ of different objective methods, for src1 to src6. The number of outliers and $RMSE^*$ values are also indicated in each plot. Error bars indicate the 95% confidence intervals for each corresponding MOS. These plots are for temporal TMO-iTMO.



(g)



(h)



(i)

Figure 3: (continued) MOS (subjective scores) and $MOS_{objective}$ of different objective methods, for src7 to src9. The number of outliers and $RMSE^*$ values are also indicated in each plot. Error bars indicate the 95% confidence intervals for each corresponding MOS. These plots are for temporal TMO-iTMO.