## 1.1 Why do we need objective video quality metrics?

It is widely accepted that objective measurement of visual quality is one critical tool in the drive to deliver better QoE when streaming video. There are many applications where the use of subjective Mean Opinion Scores (MOS) gathered experimentally with human subjects is neither practical, affordable or scalable.

The use of objective – i.e. algorithmic – visual quality metrics underpins techniques such a quality-adaptive streaming and context aware encoding, as well as QoE analysis tools such as Eurofins *Quality of Analysis Tool*. Several objective video quality algorithms are available, so which should you use?

## 1.2 Which one do others think is best?

Firstly, we need to define what we mean by "best". Typically, it is taken to mean how closely the score of the algorithm matches the quality score given by humans; or, to be more precise, the mathematical correlation between the algorithm's score and the ITU defined MOS, obtained using one of the experimental methods prescribed by the ITU. There are broadly three categories of algorithms:

- **Referenceless or no-reference metrics:** These metrics have access to the final decoded video, but no information about the original, pre-encoded video.
- **Full-reference metrics:** These metrics have complete access to both the final decoded video and original, pre-encoded video.
- **Reduced reference metrics:** This hybrid approach gives the algorithm some metadata about the pre-encoded video, but not the full bit-stream.

The latest full-reference metrics show the closest correlation to MOS evaluations. There are many cases, for example live network monitoring, where only no-reference or reduced reference metrics are practical. For our purposes, where we were designing a tool to give the best possible insight into QoE for offline, analytical purposes it was clear that a full-reference metric was the best choice.

Beyond the credibility of the score itself, there are other factors to consider, such as computational complexity and the practical usability of tooling which implements the algorithm. Various articles have compared the various objective visual quality algorithms, including the following:

1. An early comparison that pre-dates VMAF: https://ece.uwaterloo.ca/~z70wang/research/ssimplus/
2. Netflix's announcement of VMAF: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652
3. SSIMWAVE's subsequent comparison of VMAF and SSIMPLUS: https://www.ssimwave.com/download/7861/

All of these evaluations could be accused of bias. The publishers of the comparison are also the originators of an algorithm, and – surprise, surprise! – they find their own algorithm to perform best. To our knowledge there has been no independent, peer-reviewed, widely cited research as to which objective full-reference video quality algorithm is best.

## 1.3 Which one do we think is best?

It seemed clear (in early 2017) that the two standout candidates were SSIMPLUS and VMAF, so we decided to conduct our own comparison of these two algorithms. VMAF is available on Github (https://github.com/Netflix/vmaf), whereas SSIMPLUS is a commercial tool available from SSIMWAVE (https://www.ssimwave.com). Non-proprietary ancestors of SSIMPLUS, including SSIM and Multiscale SSIM, are also available and should not be confused with SSIMPLUS which has additional functionality.

The graphs below show an example of some of the comparisons we carried out using the 'Tears of Steel' video content, encoded with H.264 at 24fps and a resolution of 1920x856. For this experiment we compared a 12Mb/s and 5Mb/s encoding as we were interested in understanding whether one metric tended to be more discriminatory for high quality encodings.


Figure 1: SSIMPLUS score

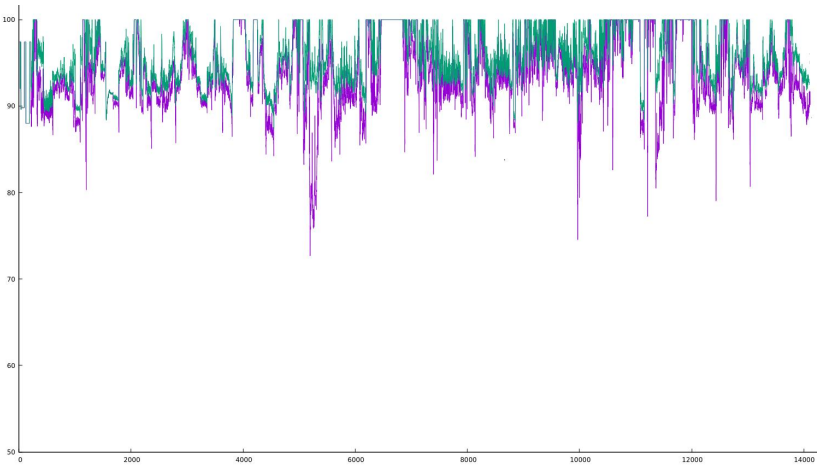


Figure 2: VMAF score

In both graphs, the horizontal axis is the frame number and the vertical axis is the metric score. The purple line represents the 5Mb/s encoding and the green line represents the 12Mb/s encoding. From this and similar comparisons we concluded that neither metric was more discriminatory than the other or performed obviously better on specific sequences where there were discernible video artefacts. Although the VMAF scores are higher on average in the above sequence, this was not true for all sequences. A clear difference is the “noisiness” of the metric and this is considered further below.

Following our evaluation, we concluded the three key advantages of SSIMPLUS for our purposes were as follows:

1. SSIMPLUS takes into account the type of display when assessing the visual quality – specifically, the size and resolution of the device. Intuitively this is obviously necessary. For example, we know that 720p video on a 60 inch 4K TV will sometimes look non-optimal even if

it is encoded with adequate bit rate, whereas the same video on a standard iPhone could achieve close to maximum perceived quality for that screen. Conversely, while it makes sense to display the highest bit rate variant available on that same 4K TV, it makes no sense for a small screen mobile device to download a variant with a much higher bit-rate that will result in very little perceptible quality improvement. Clearly, the display type matters when assessing visual quality, yet most metrics don't consider this effect which detracts from their practical usage.

The SSIMWAVE toolset for SSIMPLUS supports many pre-set display devices and SSIMWAVE will add support for new device types if requested.

2. Most visual quality algorithms assess the visual quality on a frame-by-frame basis. I.e. the score of each frame is measured entirely independently from the score of previous frames. Of course, the human visual system does not work in that way: our perception of quality does not rise and fall at, say, 25 or 50 times per second, but is subject to a hysteresis effect. The VMAF algorithm does not attempt to model this effect, so when you compare the VMAF and SSIMPLUS scores for the same video file you can see that the SSIMPLUS curve is much smoother.
SSIMWAVE describes this aspect of SSIMPLUS as follows:

*"SSIMPLUS algorithm does take the temporal elements into consideration during quality assessment. It is a hybrid model that covers many psycho-visual factors of human visual system, such as those described in* [1]. *The smoothing comes from the short-memory effect of the human brain and the time interval is content-adaptive and related to video characteristics as well, because humans tend to memorize simple content with high quality much longer than a complex one at low quality."*

[1]Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," Journal of the Optical Society of America A, vol. 24, no. 12, pp. B61-B69, Dec. 2007

3. SSIMPLUS is able to measure video quality for content that has a different resolution or frame-rate to the reference content. If you are prepared to do the video processing yourself then it is possible to compare streams of different resolution or different frame-rate using VMAF (or any other algorithm). This requires some familiarity with digital signal processing – e.g. the need for temporal and spatial low pass filtering to prevent unwanted aliasing caused by down sampling and up sampling – and can be achieved with FFMPEG. Nevertheless, for many users the need for this additional processing does hinder out-of-the-box use of other algorithms and SSIMPLUS's built in support for cross-frame rate and cross-resolution comparison is very convenient.

Having explained the motivation behind our choice of SSIMPLUS, it is important to mention that it is possible to use other objective video quality algorithm within Eurofins' automated video streaming QoE analysis framework.
To find out more about this framework and to understand the type of analysis it enables please visit http://bit.ly/QoETesting or contact DigitalTesting@eurofins.com.

## About Eurofins Digital Testing

Eurofins Digital Testing is the world's leading Digital TV Quality Assurance (QA) specialist, providing test tools, test services and training for manufacturers, standards organisations, software developers and Digital TV Operators. We operate globally with test lab facilities in the US, Europe and Asia.