Measuring the quality of video

As video becomes a fundamental part of advanced networking applications, being able to measure its quality has significant importance at all stages: starting from the development of new video codecs and ending to the monitoring the quality of the transmission system. Indeed, in order to appreciate and compare the performance of all involved components and devices, methods to assess and quantify the quality are proposed. In general, two classes of methods are available to measure video quality: subjective tests, where human subjects are asked to assess or rank the viewed material, and objective models, which are computational models that measure the quality by comparing the original and distorted video sequences. Subjective tests may produce the most accurate ratings, but, they require costly and complex setup and viewing conditions and thus, they are inflexible to use. Objective quality metrics, on the other hand, are based purely on mathematical methods, from quite simplistic yet inaccurate models, like PSNR, to sophisticated ones that exploit models of human visual perception and produce far more reliable results. While objective models of quality appear to be very promising, both methods are considered useful in the process of measuring the quality of video applications (and, in general, multimedia applications). Once successful, standard objective models of video quality are developed, subjective tests can function complementary to the evaluation purposes or they can be used to validate new objective models.

The next sections provides a review of work on subjective and objective video quality assessment, presenting their respective advantages and weak points.

1.1 Subjective video assessment

Subjective quality tests aim to capture the user's perception and understanding of quality. As we pointed out earlier, the user's perception of quality is not uni-dimensional and it
depends on many factors. Besides the quality of the viewed material per se, user
perception is also content specific, i.e., whether the video material is interesting and
intriguing or it not. It has also been recognized that what determines quality also depends
on the purpose of the interaction and the level of user's engagement. The extent to which
QoS is perceived as degraded depends upon the real-world task that the user is
performing. Furthermore, depending on the application, the quality of the background
sound is also highly important. For example, it is shown that subjective quality ratings of
the same video sequence are usually higher when accompanied by good quality sound, as
this may lower the viewers' ability to detect impairments [89]. In case of some
applications, such as multimedia conferencing, users typically require higher audio
quality relative to video quality. Perceived quality also depends on other factors like
viewing distance, display size and resolution, as well as lighting conditions [2, 71, 75]. It is
also worthwhile mentioning the distinction between image quality measured by
mathematical procedures or computational models (i.e., the degree of distortion or
difference between the original and reconstructed images) and the observed quality or
image fidelity. It appears that images with higher contrast or slightly more colourful and
saturated images appeal more to human viewers, even though, according to a strict
mathematical interpretation of distortion (e.g., ), they appear distorted in comparison to
the originals [31, 35].

1.1.1 Procedures for subjective quality evaluation

Subjective quality assessment of video still remains the most reliable means of
quantifying user perception. It is also the most efficient method to test the performance of
components, like video codecs, human vision models and objective quality assessment
metrics (section 4.1.2). This procedure, called , involves formal subjective tests where
users are asked to rate the quality using a 5-point scale, as shown in Figure 1, with quality
ratings ranging from bad to excellent.

ITU-R recommendation BT.500-10 [64] formalises this procedure by suggesting several
experimental conditions, like, viewing distance and viewing conditions (room lighting,
display features, etc.), selection of subjects and test material, assessment and data
analysis methods. Depending on what contextual factors that influence user's perception
need to be derived, three testing procedures are most commonly used: Double Stimulus
Continuous Quality Scale, Double Stimulus Impairment Scale and Single Stimulus
Continuous Quality Evaluation.
1.1.1.1 Double Stimulus Continuous Quality Scale (DSCQS).

In this method of subjective quality assessment, viewers are shown multiple sequence pairs, which consist of the original, or "reference" and the reconstructed, or "test" sequence. The sequences are relatively short in duration (8-10 seconds). The reference and test sequences are shown to the user twice in alternating fashion, the order chosen randomly. Subjects do not know in advance which is the reference sequence and which is the test sequence. They rate the material on a scale ranging from "bad" to "excellent" (Figure 4.1), and the rating has an equivalent numerical scale from 0 to 100. The difference of the two ratings is taken for further analysis. This difference removes some rating uncertainties caused by the material content and viewers' experience. The DSCQS method is preferred when the quality of the reference and test sequences is similar, otherwise subjects can easily spot the small differences in the quality of the two sequences.

1.1.1.2 Double Stimulus Impairment Scale (DSIS).

In contrast to DSCQS, in this method the reference sequence is always presented before the test sequence, and there is no need for the pair to be shown twice. Rating of impairments is again on a 5-point scale, ranging from "very annoying" to "imperceptible" (Figure 1). This method is more useful for evaluating clearly visible impairments, such as noticeable artefacts caused by encoding or transmission.

The limitation of using rather short test sequences becomes a problem when we are interested in the evaluation of digital video systems over longer timescales. These systems generate substantial quality variations that may not be uniformly distributed over the time. Both the DSCQS and DSIS methods were not designed for the quality
evaluation of video transmission over packet networks, like the Internet, because of its non-deterministic behaviour and the bursty nature of encoded video. This means that, from the user's point of view, perceived quality can vary significantly over time. The first problem is that if double stimulus (showing both the reference and test video sequences) is used for longer sequences, the time between comparable moments will be too lengthy to be rated accurately. Furthermore, it is known that humans' memory is increased for more recent stimuli, when the duration of the stimulus is increased from 10 to 30 seconds [4]. In other words, for longer sequences (over 10sec or so), the most recent parts of the sequence have a relatively greater contribution to the overall quality impression. This phenomenon, called the recency effect, is long known in psychology literature [12], and it is difficult to quantify it in subjective tests. Pearson [89] has discussed several higher-order effects that influence users' quality ratings when assessing video sequences of extended duration. What is needed here is a method able to dynamically capture user's opinion as the underlying network conditions or visual content complexity change.

1.1.1.3 Single Stimulus Continuous Quality Evaluation (SSCQE).

In order to capture temporal variations of quality, the viewers are shown a longer program (typically of 20-30 minutes duration). The reference is not presented, and viewers assess the instantaneously perceived quality in the same fashion, by continuously adjusting a side slider on the DSCQS scale (from "bad" to "excellent"). The slider can be implemented either as a hardware device [32] or as software. Instantaneous quality scores are obtained by periodically sampling the slider value usually every 1-2 seconds. In this way, differences between alternative transmission configurations can be analysed in a more informative manner. The drawback of this method is that the accuracy of user rating can be compromised by the cognitive load imposed by the task of moving the slider. As program content tends to have a significant impact on the SSCQE ratings and it becomes more difficult to compare scores from different test sequences. When a model is required to link instantaneously perceived quality to an overall quality score calculated for the whole sequence, then the non-linear influence of good or bad parts within the sequence can be expressed by pooling methods like Minkowski power weighing [30]. Despite its attractive nature, the SSCQE method also exhibits several drawbacks, the most apparent of all is the impact of the 'recency or memory effect' scale judgements. Therefore the momentary changes in quality are quite difficult to be tracked, leading to problematic stability and reliability of the derived results.

1.2 Objective metrics of video quality

We stressed earlier that although subjective procedures for measuring video quality still constitute the most reliable method for gaining insight knowledge about the performance of digital video transmission systems, the complicated and costly setup of subjective tests makes this method particularly unattractive for automating the assessment procedure. The involvement of human subjects in this process makes this approach unusable when the quality monitoring systems have to be embedded into practical processing systems. For this reason, quality metrics able to produce objectively obtained ratings present an attractive alternative.
Objective quality metrics have been the subject of research for several years. The first models were designed to work on analogue video transmission. However, the recent advent of digital manipulation and transmission of video means that video material is effected in a completely different way, leading to different types of impairments. This required a new approach in the development of quality metrics that considers the impact of encoding and transmission in a digital system.

The simplest form of measuring the quality is by calculating the distortion at the pixel level. The \textit{peak signal-to-noise} (PSNR) measures the \textit{mean squared error} (MSE) between the reference and test sequences has been extensively used by the image processing community. Due to its simplicity it is still being used. Despite it is a straightforward metric to calculate, it cannot describe distortions perceived by a complex and multi-dimensional system like the human visual system, thus, it fails to give good predictions in many cases. Recent research in image processing has focused on developing metrics that use models of the human visual system, while others, exploit properties of the compression mechanism, and assess the effect that known encoding artefacts have on perceived quality. In the following sections, we therefore first briefly discuss the various types of distortions of compressed digital video transmission and then we present a review of recent work on objective video quality metrics.