https://doi.org/10.2352/EI.2022.34.9.IQSP-318

Image quality evaluation of video conferencing solutions with realistic laboratory scenes.

R. Falcon[1], *M. Patti*[1], *S. Brochard-Garnier*[1], *G. Pacianotto Gouveia*[1], *S. Torres Acevedo*[1], *T. Bergot*[1], *R. Alarcon*[1], *C. Bomstein*[1], *H. Macudzinski*[1], *P. Maitre*[1], *L. Chanas*[1], *H. Nguyen*[1], *B. Pochon*[1], *F. Guichard*[1]
[1] *DXOMARK IMAGE LABS - Boulogne-Billancourt, France*

Abstract

Video conferencing has become extremely relevant in the world in the latest years. Traditional image and video quality evaluation techniques prove insufficient to properly assess the quality of these systems, since they often include special processing pipelines, for example, to improve face rendering. Our team proposes a suite of equipment, laboratory scenes and measurements that include realistic mannequins to simulate a more true-to-life scene, while still being able to reliably measure image quality in terms of exposure, dynamic range, color and skin tone rendering, focus, texture, and noise. These metrics are used to evaluate and compare three categories of cameras for video conference that are available on the market: external webcams, laptop integrated webcams and selfie cameras of mobile devices. Our results showed that external webcams provide a real image quality advantage over most built-in webcams in laptops but cannot match the superior image quality of tablets and smartphones selfie cameras. Our results are consistent with perceptual evaluation and allow for an objective comparison of very different systems.

Introduction

The idea of video calling has existed as a technical challenge for over a century and it became more popular in the mid to late 1990s [1] [3]. The last couple of years have been signed by the Covid-19 pandemic and video conferencing has become essential for many people around the globe forced to stay at home. We are now relying on video calls to communicate with colleagues, clients, family and friends more than ever before and the use of videoconferencing systems in online learning and teaching has become increasingly important [2]. Video calls have different challenges associated to them [5] [4]; the presence of at least one person, multiple subject distances to the camera and challenging mixed lighting conditions with different dynamic ranges. Existing researches [11] aim to automatically improve the quality of videoconferencing systems but the exhaustive evaluation of image quality is always a complex task and cannot be fully represented by traditional measurements, such as the ones done on Deadleaves or Modulation Transfer Function (MTF) charts. Combining subjective and objective image quality assessment methodologies is mandatory to perform a robust benchmark. Since the processing pipeline of video conferencing cameras may include specific features, such as face detection for Auto-Focus (AF) or Auto-Exposure (AE), building representative user cases laboratory setup for objective evaluation is extremely important to improve the accuracy of the camera benchmark. There exist a number of well known methods [14] [6] [7] to evaluate the subjective video quality [8] [9] [10]. Our methods is based on perceptual rulers of image quality attributes and, since conditions on a natural scene can vary from session to session, a direct comparison to known references is always performed to inform the analyst's evaluation.

Objective

Our main goal is to obtain reliable and meaningful scenes that allow for the evaluation and classification of cameras aimed to video conferencing under representative user cases. The measurements and their interpretation should be enough to characterize the cameras on the following general attributes: exposure, color, texture, noise, focus, and artifacts. Each of these attributes is divided into sub-attributes, for example, color is divided into Color Rendering (CR), White Balance (WB), Color Shading (CS), among others. The static and temporal capabilities of many of these sub-attributes are evaluated and an aggregate score is calculated from them. A global score is computed based on each attribute's score, and it serves as an overall quality indicator for the device under test. The challenge of evaluating these devices is related to their very different attributes, such as different resolution, sensor sizes, and fields of views (FoVs) that cause different levels of distortion, perspective deformation and subject anamorphosis.

Methodology

The image quality attributes, relevant for the user experience, branch out from the general ones described on our objectives. For example, exposure can be divided into target exposure on subject's face, dynamic range, and contrast. We built scenes with charts and elements that allow us to measure the defined attribute properties. These scenes are divided into perceptual scenes or laboratory scenes. On perceptual scenes, models act in a predetermined fashion in an especially designed room, to have a repeatable and comparable evaluation between shooting sessions. These scenes are evaluated against predefined guidelines, but are always shot alongside other known reference cameras, to be able to spot any problem not coming from the device itself. The laboratory scenes are shot only with the device under test using a subset of Analyzer, DXOMARK's solution for camera image quality testing [17]. After the scene elements are set up, framing is performed either at a fixed distance to the chart or by a predefined frame. Different lighting scenarios can be run while recording video, to measure, in addition to static attributes, the capacity of the device to adapt to different light changes and other temporal attributes. The device is set on a firm base, such as an adjustable table or a tripod. The videos are recorded using the default camera application of the operating system used and no special features

are turned on unless specific software or settings are requested.

Scene description

The scenes have been defined with the purpose of covering the most of use cases and can be categorized as follows:

- Personal: the cameras have to work at close distance (<60cm) and offer a FoV optimized to fill a large area in the frame with the user's face.
- Huddle room: the cameras have to work well at a medium distance (60cm to 3m) and offer a wide field of view, so everyone close to the camera is captured
- Conference room: the cameras have to work well at a large range of distances (60cm to >5m) and offer a wide field of view with people close to the lens but a narrower field of view with people further away.

In the following Sections we will introduce the laboratory set-ups and the perceptual scenes used in the protocol. Measurements are performed over all the laboratory scenes and are used alongside the perceptual evaluation to aggregate a score for each of the tested image quality attribute.

Laboratory scenes

Laboratory scenes are carefully setup by DXOMARK's technical team to guarantee high repeatability between shooting sessions, to be able to fairly test and compare different devices. Five scenes contain a single test chart and three scenes contain at least one realistic mannequin. Scenes' frames and their respective names are shown in Figure 1 and Figure 2. The used lighting

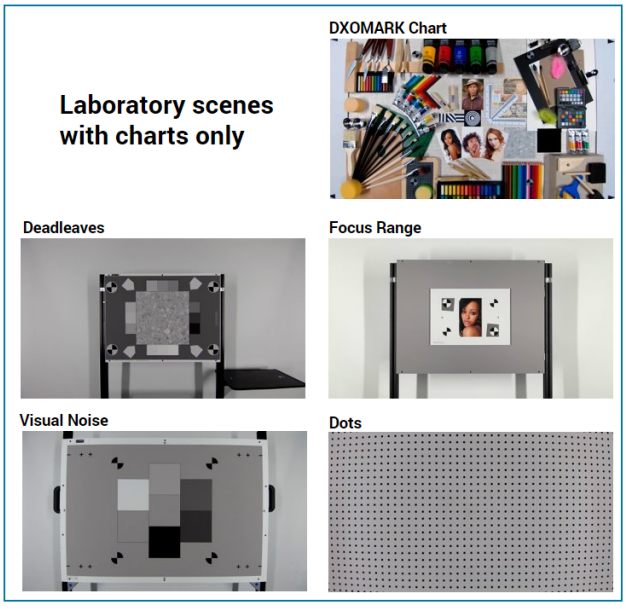


Figure 1. *Laboratory scenes that include a single test chart.*

systems in the DXOMARK laboratories are listed in Table 1 and they are systematically tested to guarantee the color temperature of the lamps used. The illuminance levels of each scene are calibrated in the center of the image frame before every session, and the lights are set to provide a uniform illumination on the region of interest of the scene.

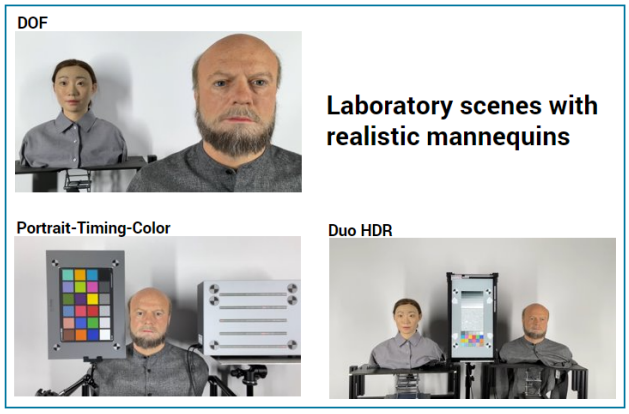


Figure 2. *Laboratory scenes that include at least one realistic mannequin.*

Laboratory lighting conditions

Name	Color temperature	Reference
Light Panel	2700K to 6500K	KinoFlo Celeb 250
LED	2700K	Philips MASTERSpot GU10 5.5W 927 25D
TL83	3000K	Philips MASTER TL5 HO 49W/830
TL84	4000K	Philips MASTER TL5 HO 49W/840
D65	6500K	Philips MASTER TL5 HO 49W/865

Three realistic mannequin heads can be used, a fair-skin male, a dark-skin female and an Asian fair-skin female (Figure 3), on future references we will be calling them Eugene, Diana and Sienna, respectively. On the mannequin, we perform a lightness

Figure 3. *Dark-skin (Diana), Asian-skin (Sienna) and fair-skin (Eugene) realistic mannequins.*

measurement to evaluate target exposure and we use an Artificial Intelligence (AI) algorithm trained over hundreds of labelled images to estimate the details preservation. The skin tone of the mannequins is an accurate metamerism with real skin, but it does not possess the exact power spectral density of human skin, and it cannot be used to evaluate the color and white balance accuracy of the device. For this reason, we added the ColorChecker® in

the presence of a face. We compare the results of these laboratory scenes with perceptual ones, shot with multiple devices, to validate their relevance to characterize image quality in a laboratory setting. For each measured metric, we developed a series of thresholds, based on user preference data, that define acceptable quality levels and assign them a quality score. The detailed measurements and evaluation methodology performed on laboratory scenes are described in Section "Laboratory measurements".

Perceptual scenes

Perceptual scenes are shot to reproduce real-life use cases that cannot be reproduced in the lab. Each scene is staged to have natural movements from the people in them, while being similar between shooting sessions. Each time a new device is tested, at least two known devices are shot simultaneously, to provide a point of reference to the perceptual analysis. These reference devices are often chosen to be of similar characteristics of the Device Under Test (DUT) in terms of price range, FoV and the use cases they claim to cover. Each scene includes a deep skin person and a fair skin person, each of them may change between shooting sessions, but are normally the same for a given device under test and its references.

Duo Backlit

This scene reproduces a back-lit conference room, on challenging lighting conditions. The composition of the scene allows for evaluation of white balance, skin tone rendering, bright and dark area preservation, as well as noise and detail preservation on the subject. Halfway through the video, a deep skin tone model arrives, challenging the AE to properly expose the room and the two individuals, the Auto-WB (AWB) to keep the right colors on the scene and the AF stability. The movement on the scene may also cause local losses of texture due to compression, increase noise on moving objects or show processing artifacts, such as ghosting. This scene is framed on the horizontal FoV, using markers on the walls of the room.

Storyboard	
	
Fair model looks at the camera and their laptop (~15s)	Deep model arrives and sits down (~5s)
	
Fair model looks away, deep model waves (~5s)	Both models look at the camera (~10s)
Elements	
Shot to frame. Fair and deep model on the same plane. Mixed lighting: Philips Master TL5 HE21W/840 and sunlight.	

Figure 4. *Duo Backlit scene storyboard and shooting conditions.*

Large room

This scene reproduces a large conference room, on typical conditions. It is used to evaluate external webcams and it is replaced by "single conference room" scene in the case of laptop integrated webcams or selfie cameras. The composition of the scene allows for evaluation of white balance, skin tone rendering, bright and dark area preservation, as well as depth of field and detail preservation on the subjects. Throughout the video, the fair model enters the scene and walks toward a screen showing the following chart. On the chart we can evaluate attributes, such as legibility, color rendering, and bright clipping. This scene is framed at fixed distances. The camera is aligned with the center of the table and the monitor, the distance between the camera and the monitor is 4m and the deep model sits 1.25m from the camera. Since the dynamic range of the scene may vary with the FoV, at least one reference with a similar FoV is shot.

120° DFoV, 90° DFoV 75° DFoV

Storyboard	
	
Deep model looks at the camera (~5s)	Fair model enters and walks to the screen (~5s)
	
Deep model looks away, fair model presents (~10s)	Both models look at the camera (~5s)
Elements	
Shot by distance. Fair and deep model on different planes. Mixed lighting: Philips Master TL5 HE21W/840, LED 22W/830 and sunlight.	

Figure 5. *Large Room scene storyboard and shooting conditions.*

Single Conference Room

The scene represents a small conference room, where users tend to be close to the camera. The model moves the head challenging the AE and AWB that are based on face detection algorithms. The movement on the scene may also cause local losses of texture due to compression, increase noise on moving objects or show processing artifacts, such as ghosting. Additional to exposure, skin tone rendering, details and noise on the subjects, other elements on the scene help evaluate quality on non-human subjects: the painting on the wall helps with color rendering and detail preservation. The whiteboard behind is scribbled with text and graphs of different colors, to evaluate legibility and color rendering of fine details. This scene is framed on the horizontal FoV,

using the painting on the walls of the room.

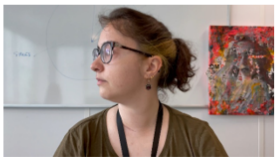
Storyboard	
	
Model looks at the camera (~5s)	Model moves the head (~10s)
	
Model waves (~5s)	
Elements	
Shot to frame Mixed lighting: Philips Master TL5 HE21W/840 and sunlight.	

Figure 6. *Single conference room scene storyboard and shooting conditions.*

Small room

The scene represents a small conference room, where users tend to be close to each other. The deep and the fair models are separated by 60cm on the horizontal axis and 60cm on the optical axis. Additional to exposure, skin tone rendering, details and noise on the subjects, other elements on the scene help evaluate quality on non-human subjects: The painting on the wall and the cards on the table help with color rendering and detail preservation. The whiteboard behind is scribbled with text and graphs of different colors, to evaluate legibility and color rendering of fine details. This scene is framed on the horizontal FoV, using markers on the walls of the room. For devices with narrow FoVs, the fair model may lean on the table, so that their face is fully visible.

Sofa

The scene represents a living room during an informal video call. The saturated colors of the wall and the sofa challenge white balance and color rendering. The position of the models and the sofa will make anamorphosis and other perspective deformations evident. This scene is framed on the horizontal FoV, using markers on the left wall and the right door of the room. Both heads of the models should be within the central horizontal thirds.

Laboratory measurements

Measurements are performed on different elements of all the laboratory scenes. The measurement values are then transformed into a score by comparing them with a certain specification value; if the measurement is within an acceptable range, the score is the maximum possible for that condition. Out that range, the score decreases until it reaches a failure point, beyond that, the score is the minimum for that condition. Laboratory scenes are classified by the illuminance E_V level used to shoot them: bright light, indoor and low light, these separations are based on the illumination

Storyboard	
	
Both models look at the camera (~5s)	Both models converse (~10s).
	
Both models look at the camera (~5s)	
Elements	
Shot to frame Fair and deep model on different planes Mixed lighting: Philips Master TL5 HE21W/840 and sunlight.	

Figure 7. *Small Room scene storyboard and shooting conditions.*

Storyboard	
	
Fair model looks at the camera (~5s)	Deep model sits down and waves (~5s)
	
Both models look at the camera (~5s)	Deep model leaves (~5s)
Elements	
Shot to frame Fair and deep model on the same plane Mixed lighting: Philips Master TL5 HE21W/840, LED 22W/830 and sunlight.	

Figure 8. *Sofa scene storyboard and shooting conditions.*

conditions typical to those use cases:

- Bright light: $E_V \geq 500lux$
- Indoor: $500lux > E_V \geq 100lux$
- Low light: $100lux > E_V$

Exposure

The exposure attribute is measured on Portrait Timing-Color and Duo HDR (Figure 2). The evaluated metrics are static target exposure, temporal exposure, and entropy. The target exposure of a region of interest (ROI), is measured on the mannequin's forehead as the average Lightness (L*) in the CIELAB (L*a*b*) color space. L*=0 yields black whereas L*=100 is white. The temporal exposure is evaluated by 4 metrics on the ColorChecker®: end lightness, overshoot, oscillation time and convergence time. The Figure 9 explains the four concepts. Entropy is a single-value metric computed from the composite HDR chart (Duo HDR) that

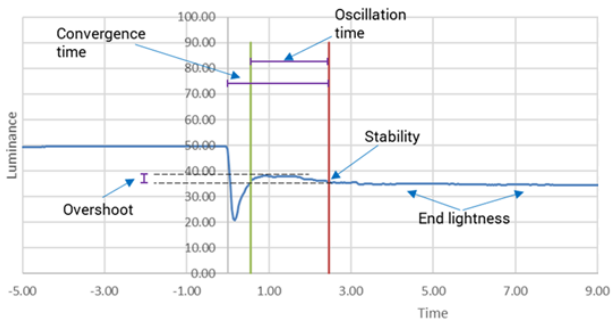


Figure 9. *Illustration of convergence time, oscillation time, overshoot, and stability on a transition from 200lux to 26lux.*

measures the local contrast preservation and represents the quantity of information contained in it. The HDR chart is set to have +2EV and +4EV illuminance difference with Eugene's face. The value for entropy is given in bits, and the maximum possible value for an 8-bit video camera is 8 bits. The entropy of a given crop of an image is given by computing its histogram and using the formula:

$$\text{Entropy} = \sum_{K} hist\,(K)\, log_2\left(\frac{1}{hist\,(K)}\right) \tag{1}$$

where *hist* is the histogram of the gray levels and K the gray level (from 0 to 255). In the context of the HDR measure, entropy is computed using a grayscale of 63 gray patches.

Color

The color attribute is measured on the Portrait-Timing-Color scene and its attributes are Color Rendering (CR), Static White Balance (SWB), Temporal Color Rendering (TCR) and Temporal White Balance (TWB). Color rendering and white balance are calculated in the L*a*b* space, and are defined as the difference between the color coordinates a^*_S,and b^*_S measured on the ColorChecker® of the video, and the reference values a^*_R,and b^*_R, using the formula:

$$\Delta ab^* = \sqrt{\left(a^*_R - a^*_S\right)^2 + \left(b^*_R - b^*_S\right)^2} \tag{2}$$

The reference values consider the D65 lighting condition and a scoring system, defined by acceptable ranges in the L*a*b* space, takes into account all the different lighting conditions. The magnitude of the metric Δab^* is not enough to quantify if the color rendering or white balance will be pleasant to a user, the angle of the deviation (representing the hue shift from the nominal color) is as important. For example, a white balance Δab^* towards the yellow will give a pleasant, warm, feeling to the scene, but a similar deviation towards pink will feel unnatural. Our scoring system takes this into account, defining the acceptance range and the failure points as ellipses in the L*a*b* space. These ellipses are specific to each color of the ColorChecker® when evaluating color rendering, and to each illuminant color temperature when evaluating white balance.
Although a score is measured for each patch, at each lighting condition, on the report an aggregated CR metric is used to summarize the measurements, for each patch n.

$$\text{CR} = \frac{1}{n}\sum_{0}^{17} \Delta ab^*_n \tag{3}$$

A similar metric is calculated for white balance on the gray patches.

$$\text{WB} = \frac{1}{n}\sum_{19}^{22} \Delta ab^*_n \tag{4}$$

The patches 18 (white) and 23 (black) are not used for the white balance computation. The Temporal Color Rendering and Temporal White Balance are evaluated as the oscillation intensity, which is computed as the integral difference between a short-term moving average (0.25 s) and a long term moving average (1 s) over a ramp of continuous change of illuminant [t_{end} , t_{begin}].

$$\text{E}_{short}(t) = 4\int_{t-0.25}^{t} \Delta ab^*(\tau)\, d\tau \tag{5}$$

$$\text{E}_{long}(t) = \int_{t-1}^{t} \Delta ab^*(\tau)\, d\tau \tag{6}$$

$$\text{I} = \int_{t_{begin}}^{t_{end}} ||\,\text{E}_{short}(t) - \text{E}_{long}(t)\,||\, dt \tag{7}$$

Focus

The focus attribute is measured on the Focus Range scene. The evaluated metric is focus range, which indicates the distances for which the device is capable to produce a sharp image. The depth of field, that contributes to the Focus score, is perceptually evaluated on the DOF scene (Figure 2). The focus range is evaluated by calculating the edge acutance of the chart. The acutance is a single value metric calculated from a MTF result. The value is a weighted average of the MTF, with weights dependent on the CSF (contrast sensitivity function) of the typical human eye and the viewing conditions (image size and distance to subject). The metric therefore only takes into account the visible frequencies for a given viewing condition. The acutance can be computed as:

$$\text{A} = \frac{1}{A_r}\int_0^{\infty} MTF_p^{-1}(\nu)\, MTF(\nu)\, MTF_d(\nu)\, CSF(\nu)\, d\nu \tag{8}$$

With $A_r = \int_0^{\inf} CSF(\nu)\, d\nu$, $CSF(\nu)$ is the contrast sensitivity function of the eye at a given distance from the viewing screen, $MTF_p(\nu)$ is given by the resolution of the print of the chart itself, $MTF(\nu)$ is the measured MTF and $MTF_d(\nu)$ is given by the resolution at which the video is being watched. For each device to chart distance captured, 10 frames are extracted from the video, their edge acutance is measured, and the values are averaged into A_d, the acutance for that distance. The focus range score is calculated from the standard deviation of the A_d

$$\sigma_A = \sqrt{\frac{\sum (A_d - \overline{A}_d)^2}{N}} \tag{9}$$

Since an important acutance drop would increase σ_A, indicating a loss of focus.

Texture

The texture attribute is measured on the Portrait Timing-Color, the Deadleaves and the DMC scenes. Its attributes are detail preservation, texture acutance and edge acutance. The detail preservation is estimated by comparing the detail level of the Fair Mannequin head and the DMC against a ruler of annotated images. At the current revision of this document, the scoring on the DMC is done by a trained AI algorithm [12] and has a score scale from 0, the worst, to 100, the best observed. The scoring on the fair mannequin head is done perceptually, and the results are being used to train an AI algorithm for this scene [13], and has a score scale from 0, the worst, to 100, the best observed. The edge acutance metric on the Deadleaves chart is identical to the one used on the focus range measurement. The texture acutance metric uses the same calculation for the acutance, but estimates the MTF of the optical system with the Deadleaves portion of the chart. The statistics of this model follow the distribution of the same statistics in natural images. The use of this model for sharpness measurement has been described by Cao et al. [15]. This method is less susceptible to over-sharpening than the slanted edge measurement used for the edge acutance measurement. The texture acutance and the edge acutance computed from the Deadleaves chart do not contribute to the final texture score.

Noise

The noise attribute is measured on the visual noise scene. Its attributes are visual noise, temporal noise, and noise chromaticity A typical measurement is the signal to noise ratio (SNR) calculated on a flat area of the image. This metric does not account for viewing conditions or the sensitivity of the eye, so a spatial filtering using the $CSF(\nu)$, is used for the DXOMARK visual noise measurement. Since noise depends on luminance, these variances are computed on seven ROIs with different reflectances. A normalization is required to compare different devices with different exposures: the noise is given for CIELab L*= 50, linearly interpolated from the two closest uniform grey patches to lightness value L*= 50. Once the spatial filtering and the normalization is done, the DXOMARK Image Labs visual noise is the base-10 logarithm of the weighted sum of the L*,a*,b* variances.

$$\Omega = K\, log_{10}\left[1+\sigma^2(L^*)+\sigma^2(a^*)+\sigma^2(b^*)\right] \tag{10}$$

The Temporal Visual Noise metric is computed in the CIELab color space. Exposure and white balance are compensated to isolate the temporal noise from other distinct effects (exposure drift, white-balance drift, etc.). We measure temporal noise variances $\sigma^2_{L^*}$, $\sigma^2_{a^*}$ and $\sigma^2_{b^*}$ as spatial averages (over the pixels of an entire patch) of the temporal variances (over all frames for a single pixel). A lightness normalization is performed as for Spatial Noise. The perception of temporal noise shows a good correlation to the square root of a weighted sum of the noise variances. Thus, we set all weights to 1, and we define the temporal visual noise (TΩ) as:

$$\mathrm{T}\Omega = \sqrt{\sigma^2_{L^*}+\sigma^2_{a^*}+\sigma^2_{b^*}} \tag{11}$$

Notice that this is the average euclidean distance of each pixel of a correctly exposed patch from its average.

Artifacts

The artifacts attribute is measured on the Dots, Deadleaves and Portrait Timing-Color scenes (see Section . Its attributes are Lens Geometric Distortion (LGD), Lateral Chromatic Aberration(LCA), Ringing Intensity (RI) and Frame Rate (FR). The LGD and LCA are measured using the Camera Phone Image Quality (CPIQ) LGD and LCA metric on the Dots chart as described on IEEE's Camera phone Image Quality [16] The RI measurement is performed on the slanted edges of the Deadleaves chart. Measurements are given for each color channel, in the same zones in both directions. The method consists in extracting the exact profile of the edge transition with sub-pixel accuracy. The maximum of ringing oscillations is measured in both dark and bright zones. Its amplitude is given as a percentage of the step edge amplitude. The FR is evaluated on the LED Universal Timer (Portrait-Timing-Color scene) in different illumination conditions. The measurement is performed by measuring how many frames it takes the bright LED in the box to complete a revolution. Since the revolutions per second of the LED are known, the Frame rate can be calculated.

Video conferencing benchmark

At the time of writing, we tested and scored a pilot database of six external video conference webcams, five laptop integrated webcams and four selfie cameras of mobile devices. For each category, we have picked cameras that cover a wide price range and are targeted to business use as well as personal use. External video conference webcams are the most advanced cameras, designed to work in huddle and conference rooms with groups of people and offer intelligent framing and focus features as well as motorized tilting and panning. Integrated webcams in laptops and selfie cameras of tablets and smartphones are designed to work mostly at close distance (<60cm) and offer a FoV optimized so the user's face fills a large area in the frame. To create a level playing field, all external webcam were tested using the same recording app, laptop and selfie cameras were tested using the native app of their operative system. All the cameras were tested at the maximum resolution of the device, capped at 1080p, and with the same maximum frame rate (30fps). While higher resolutions and frame rates are available on some models for smooth video, they require internet speeds that are often not available in homes or offices. All other settings remained at their default "out-of-the-box" value. We did not test any special features, manual settings or pan-tilt functions. All cameras used the latest available firmware at the time of testing. The criteria for testing videoconferencing cameras are somewhat different to those for testing smartphone cameras or DSLRs. The most important image quality attributes for videoconferencing are exposure and color. If any of those two go wrong they have a direct detrimental effect on the entire experience. An inaccurate focus or geometric artifacts, such as distortion, can have a negative impact, too, but are less important in comparison. Texture and noise, which are crucial to photographers who print their images or display them at large size, are of less importance as they may be reduced by video compression or network issues. In terms of light conditions, the highest weight is given to the indoor light levels and sources as most videoconferences take place indoors. The perceptual score of each image quality attribute is derived both from rulers and from a direct comparison to known references shot simultaneously. This is always

performed for the perceptual scenes in order to consider the variable conditions on a natural scene between different test sessions. The next Section summarizes the average scores and measurements divided by camera category.

Results

The results showed that external webcams provide a real image quality advantage over most built-in webcams in laptops. In addition they come with more features that are not yet covered by our test protocol. Things looks slightly different when making the comparison to selfie cameras of premium tablets and ultra-premium smartphones, though. These devices are capable of delivering excellent image quality, thanks to a combination of powerful hardware, computational imaging and dedicated image processing pipelines.

Figure 10 shows the results of the target exposure measured on the mannequin's face in HDR conditions (FaceHDR) and SDR conditions (FaceSDR) as described in Section "Exposure". A L* value within the range of 55 and 75 is considered as acceptable. The entropy, as measure of dynamic range (DR) in HDR conditions, is also shown on the radar chart. An entropy value less than 6 bits is indicative of exposure clipping. The values of the radar chart are average measures of tested conditions as follows:

- **HDR** : 1000lux D65 (+4EV, +2EV); 100lux TL84 (+4EV, +2EV); 20lux LED (+4EV)
- **SDR** : 1000lux D65; 300lux TL84; 100lux TL84; 20lx LED; 5lux LED

The values of the bar chart are exposure scores resulting from the perceptual analysis on perceptual scenes. In our image quality

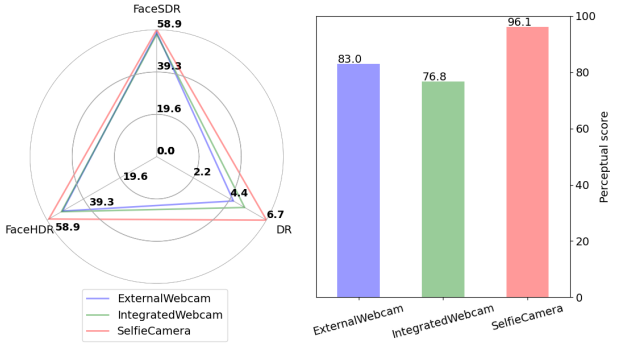


Figure 10. *Exposure scores. The radar chart shows the average L* measured on the mannequin's face in HDR and SDR condition. DR is the average entropy measured in HDR conditions. Perceptual scores are shown on the bar chart.*

tests all cameras deliver acceptable face exposure in Standard Dynamic Range (SDR) condition. They do however differ in terms of Dynamic Range (DR) and the handling of face exposure in difficult back-lit scenes. The selfie cameras provide a wider DR allowing a more accurate face exposure in HDR conditions. The integrated webcams have a slight advantage over the external webcams in DR performance. However, built-in webcams in laptops are prone to temporal exposure instabilities that affect the perceptual score.

Figure 11 shows the results of the CR, SWB and TWB measured on the ColorChecker® as described in Section "Color". The values of the radar chart are averaged scores of different lighting scenarios with multiple transitions. An example of used scenario is shown in Figure 12. The values of the bar chart are color scores resulting from the perceptual analysis on perceptual scenes. The external webcams provide a more stable and accurate WB rendering but their CR accuracy is the worst of the three categories. The selfie cameras provide the best CR accuracy and a pleasant WB rendering placing them above the external webcams in perceptual score. The laptop integrated webcams often show inaccurate WB casts and temporal instabilities.

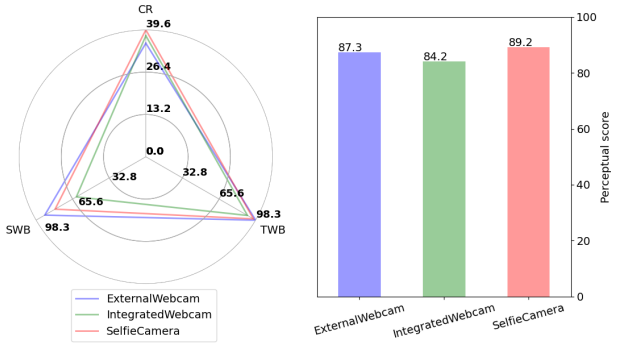


Figure 11. *Color scores. The radar chart shows the average CR, TWB and SWB scores as result of measurement on the ColorChecker® under lighting scenarios with multiple transitions. Values are expressed in in the L*a*b* space units and converted into scores by considering ranges of acceptance within the L*a*b* space. Perceptual scores are shown on the bar chart.*

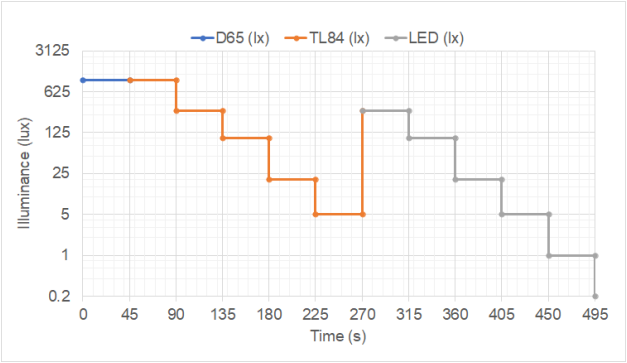


Figure 12. *Example of lighting scenario used during video recording.*

Figure 13 shows the results of details preservation measured by the AI algorithm on the DMC and mannequin face as described in Section "Texture". The values of the radar chart are averaged measures of bright light (1000lux D65), indoor (300lux TL84; 100lux TL84) and low light (20lx LED; 5lux LED) illuminance conditions. The percentage of measured ringing artifact is also shown as indication of texture over-sharpening. The values of the bar chart are texture scores resulting from the perceptual analysis on perceptual scenes. The external webcams provide the best detail preservation in bright light and indoor conditions but they tend to over-sharp the video. The selfie cameras provide well-balanced detail preservation among all the illuminace conditions and the lower ringing intensity, with respect to external webcams,

allows a more natural detail rendering resulting in a better perceptual score. Built-in webcams in laptops are limited by sensor size and number of pixels providing the worst texture quality of the tested categories.

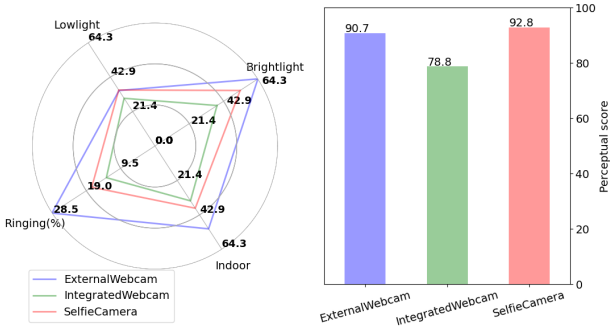


Figure 13. *Texture scores. The radar chart shows the average details preservation measured by the AI algorithm on the DMC and mannequin face under bright light, indoor and low light illuminance conditions. The percentage of ringing artifact measured on the Deadleaves chart is an indication of texture over-sharpening. Perceptual scores are shown on the bar chart.*

Figure 14 shows the results of noise measured on the visual noise chart as described in Section "Noise". The values of the radar chart are averaged measures of bright light, indoor and low light illuminance conditions. Video of visual noise chart is recorded by using a single lighting scenario shown in Figure 12. Spatial and temporal component of the noise are also shown as average measure of all tested conditions. The values of the bar chart are noise

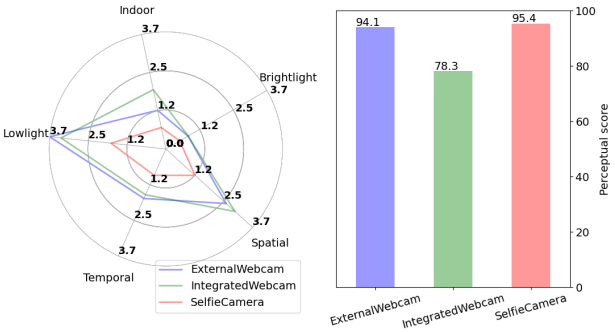


Figure 14. *Noise scores. The radar chart shows the average visual noise measurements under bright light, indoor and low light illuminance conditions of Figure 12 scenario. Spatial and temporal component of the noise are also shown as average measure of all tested conditions. Perceptual scores are shown on the bar chart.*

scores resulting from the perceptual analysis on perceptual scenes. The selfie cameras provide the lowest level of noise thanks to efficient de-noising algorithms achieving also the highest perceptual score. The laptop integrated webcams present a relative high level of noise in indoor conditions where all our perceptual scenes (as most videoconferences) take place. This measure is confirmed by the lowest perceptual noise score among the camera categories. It is worth mentioning that the highest noise component is spatial and it is due to a combination of limited pixels size and temporal de-noising algorithms.

Figure 15 shows the results of focus measured on the focus range chart as described in Section "Focus". Objective score is the maximum for external webcams since they are able to focus the target at every tested distance. However, their perceptual score is affected by the auto-focus instabilities that are sometimes visible in presence of moving content in the scene. Selfie cameras and laptop integrated webcams do not achieve a perfect focus objective score due to a significant drop of acutance when the target distance is about 30cm. Some of the selfie cameras and integrated webcam also suffer of a relative narrow depth of field that affect the perceptual focus score when multiple subjects at different distances are present in the scene.

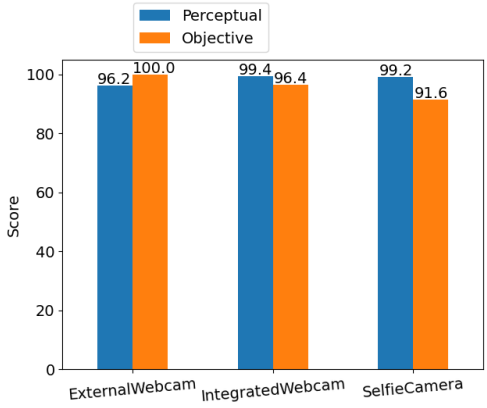


Figure 15. *Focus objective and perceptual scores.*

Conclusion

In this paper we propose a novel methodology to assess the image quality of cameras for video conferencing. Realistic mannequins are used as an alternative to charts to perform IQ measurements such as target exposure and details preservation based on AI. We perform a complete characterization and benchmarking of video conference cameras through image quality analysis that correlates the measurement results from the objective methods with subjective results in real user experiences framework.

References

[1] Sondak, N.E. and Sondak, E.M. "Video Conferencing: The Next Wave for International Business Communication". Annual Conference on Languages and Communication for World Business and the Professions (1995).

[2] Correia, Ana-Paula, Chenxi Liu, and Fan Xu. "Evaluating videoconferencing systems for the quality of the educational experience." Distance Education 41.4 (2020): 429-452.

[3] Tang, John C., and Ellen Isaacs. "Why do users like video?." Computer Supported Cooperative Work (CSCW) 1.3 (1992): 163-196.

[4] Gough, Michael. Video conferencing over IP: Configure, secure, and troubleshoot. Elsevier, (2006).

[5] Mouzourakis, Panayotis. "Videoconferencing: Techniques and challenges", Interpreting 1.1 (1996): 21-38.

[6] S. Winkler, "Video Quality Measurement Standards – Current Status and Trends", 7th International Conference on Information, Communications and Signal Processing (ICICS). pg. 1-5. IEEE, (2009).

[7] J. You, U. Reiter, Miska M. Hannuksela, M. Gabbouj, A. Perkis,

"Perceptual-based quality assessment for audio-visual services: A survey", Signal Processing: Image Communication 25.7 (2010): 482-501.
[8] ITU-R Recommendation BT.500-12, "Methodology for the subjective assessment of the quality of television pictures", 2009.
[9] ITU-R Recommendation BT.710-4, "Subjective assessment methods for image quality in high-definition television", 1998.
[10] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications", 2008.
[11] Shi, Cuizhu, et al. "Automatic image quality improvement for videoconferencing." 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 3. pg. iii-701 IEEE, (2004).
[12] Tworski, Marcelin, et al. "DR2S: Deep Regression with Region Selection for Camera Quality Evaluation." 2020 25th International Conference on Pattern Recognition (ICPR). pg. 6173-6180. IEEE, (2021).
[13] Nicolas, Chahine, and Belkarfa Salim. "Portrait Quality Assessment using Multi-Scale CNN." London Imaging Meeting. Vol. 2021, pg. 5-10. Society for Imaging Science and Technology, (2021).
[14] Berndtsson, Gunilla, Mats Folkesson, and Valentin Kulyk. "Subjective quality assessment of video conferences and telemeetings." 2012 19th International Packet Video Workshop (PV). pg. 25-30. IEEE, (2012).
[15] Cao, Frédéric, Frederic Guichard, and Hervé Hornung. "Measuring texture sharpness of a digital camera." Digital Photography V., Vol. 7250, pg. 7250H. International Society for Optics and Photonics, (2009).
[16] BOARD, Corporate Advisory. IEEE Standard for Camera Phone Image Quality.
[17] DXOMARK Analyzer, https://corp.dxomark.com/analyzer/

Author Biography

Rafael Falcon is Product Owner at DXOMARK. He has worked with the ISP tuning teams of many of the largest photo and video camera manufacturers, from DSLRs to smartphones, offering technical solutions to achieve the best quality, while always considering the final user experience. His responsibilities include team management, design of new photo and video quality benchmark protocols, laboratory tests, definition of shooting plans and training of IQ engineers of DXOMARK and other companies.