https://doi.org/10.2352/ISSN.2470-1173.2020.15.COLOR-162

Comparing a spatial extension of IC$_T$C$_P$ color representation with S-CIELAB and other recent color metrics for HDR and WCG quality assessment

Anustup Choudhury, Scott Daly; Dolby Laboratories Inc.; Sunnyvale, CA, USA

Abstract

Content created in High Dynamic Range (HDR) and Wide Color Gamut (WCG) is becoming more ubiquitous, driving the need for reliable tools for evaluating the quality across the imaging ecosystem. One of the simplest techniques to measure the quality of any video system is to measure the color errors. The traditional color difference metrics such as ΔE_{00} and the newer HDR specific metrics such as ΔE_Z and ΔE_{ITP} compute color difference on a pixel-by-pixel basis which do not account for the spatial effects (optical) and active processing (neural) done by the human visual system. In this work, we improve upon the per-pixel ΔE_{ITP} color difference metric by performing a spatial extension similar to what was done during the design of S-CIELAB. We quantified the performance using four standard evaluation procedures on four publicly available HDR and WCG image databases and found that the proposed metric results in a marked improvement with subjective scores over existing per-pixel color difference metrics.

Introduction

Millions of devices now support High Dynamic Range (HDR) and Wide Color Gamut (WCG) content. Display design, video processing algorithm design, system format development and comparison across products all require being able to evaluate the quality of HDR images in a perceptually relevant manner. There is a vast body of literature on quality metrics relating to image data compression for traditional standard dynamic range (SDR) images and a steadily increasing amount of research on HDR [1], but these are almost exclusively based exclusively on the luminance component. However, color distortions are also very important to assess because they are increasingly likely in HDR[1] systems, due to the additional gamut and tone mapping operations that are required to convert the source color volume to a typically reduced display color volume. Most commonly used color difference metrics have been designed to measure differences between simple test patches, as opposed to natural (complex[2]) imagery. Typically, the size of the test patch does not match the size of the objects in an image, which are often substantially smaller in terms of visual degrees (by a factor of 1/300 in consumer TV applications). Other key differences between test patches and complex imagery is that test patches typically consist of lower spatial frequencies and non-contiguous regions, while the natural imagery has much higher frequencies, masking due to textures, as well contiguous color region effects and gradients. Furthermore, while there has been significant evaluation for color differences in complex (natural and civilized) imagery [2, 3], most of this evaluation has been dominated by still images as opposed to video, and has been almost entirely limited to Standard Dynamic Range (SDR) content.

For this work, we evaluated several color difference metrics on four publicly available HDR databases consisting of complex images with various distortions, along with the corresponding subjective scores. The different databases focus on different distortions and the aggregation of these covers a wide variety of both luminance and chromatic distortions. Some of these distortions result from tone-mapping and gamut mapping operations, which tend to be dominated by lower frequencies due to shallow gradients and easily-visible regions, but also can contain step edge artifacts (i.e., having a 1/f spectrum). Other distortions include higher frequency distortions resulting from compression artifacts by various compression schemes such as JPEG, JPEG-XT, JPEG2000, and HEVC. These typically include ringing around sharp edges (‘mosquito noise’) and visible transform block boundaries (‘blocking’). While perceptually dominated by luminance distortions, these also contain chromatic distortions due to chroma subsampling and different processes acting on Y, Cr and Cb signals. Image statistics play a key role in imaging product design, as the era of displays being able to expect a certain class of imagery (e.g., optically-captured, or text, or computer-generated) has given way to systems being used for all types of imagery. Some key aspects of image statistics have been studies for SDR, such as the $1/f^N$ spatial frequency power spectra, the log-normal luminance histograms, and the principal components in descending variances of an achromatic and two uncorrelated chromatic components. However, the same statistics for HDR images are less well understood, and even for SDR these statistics do not take into account characteristics such as as texture vs. smooth gradients, mixed illumination, frequency of emissive light sources, depth of field, etc [4]. Consequently, to allow for robustness, it is desirable to evaluate as many images as possible. Toward that goal, this work includes a total of 46 source images and a total of 532 distorted images as evaluated by 94 observers across all four databases.

In addition to the commonly used (ΔE_{00}) color difference metric, we compare several recent metrics derived for HDR applications: ΔE_Z based on the J$_z$a$_z$b$_z$ color space, and ΔE_{ITP} based on the IC$_T$C$_P$ color space. While the CIE L*a*b*-based metrics have been shown to perform well for many SDR applications

[1] Since most HDR systems are also WCG, we will use the term ‘HDR’ to include both types of advances

[2] There are specific categories of imagery such as natural (i.e., no human-made objects), civilized (including human-made objects), real-world (optically captured), synthetic (computer generated) so in this paper we will use the term ‘complex imagery’ to include all cases

(product surface colors, graphic arts printing), they are known to have significant problems with new image characteristics enabled by HDR, such as shadow detail spanning several log units, emissive colors, specular reflections, scenes of mixed illumination, and interscene adaptations (e.g., from scene changes like going into a cave, or turning on a light source). One source of the CIE L*a*b* problems is the cube-root based nonlinearity forming the backbone of L*, since it has the inability to handle the continuously differing non-linearities such as log behavior (Weber's Law) for greater than $200cd/m^2$, the square-root behavior (Rose-Devries law) for less than $0.001cd/m^2$, and the continuum of changes between these extremes. On the other hand, CIE L*a*b* is vetted by standards and has a long history of use, with many experts familiar with how to apply it. While we have omitted ΔE_{94} in this analysis, it was found to perform worse than ΔE_{00} in a HDR-WCG study [5].

Both of the new HDR metrics should have better performance over the larger luminance ranges common to HDR, as they have been derived from the behavior of the CSF (contrast sensitivity function $\neq$ spatial MTF) across a large range of light adaptation [6, 7], whereas the L* essentially models a single fixed adaptation level. In addition, they are both physiologically more realistic models since they apply a nonlinearity to the signals for the actual known L, M, S, cones in the retina, as opposed to the psychophysically-derived X, Y, Z color matching signals as do the CIE L*a*b*-based metrics (which were developed before the known L, M, S responses were measured). Working with LMS cone responses allows for chromatic adaptation to be modeled without the well-known 'Wrong Von-Kries Adaptation' distortions that occur when working on the XYZ signals [8]. In particular, the metric ΔE_{ITP} is derived from IC_TC_P color space, which models the case for a hull under variable light adaptation [9]. Rigorous testing in an experimental design that allowed for short-term chromatic adaptation around a D65 bias point, and longer-term luminance adaptation was used to fine-tune the matrix used to transform from the nonlinear L, M, S responses to the opponent colors [10]. While the $J_za_zb_z$ color space shares the achromatic signal of IC_TC_P, it forms the opponent color channels differently, with a key difference being a matrix model of asymmetric cone-cross-coupling before the non-linearity is applied.

In developing color spaces, there is always the issue of whether to design it based on detection or appearance. 'Detection' focuses on small visible differences and can characterize threshold behavior. 'Appearance' addresses supra-threshold differences, and is primarily concerned with perceptual 'lengths' (i.e., particularly the larger lengths or distances across the color space). It is known that a single non-linearity cannot describe both detection and appearance effects [11, 12]. In addition, it is known that a color space designed to achieve hue linearity cannot achieve uniform detection [13]. The term micro-uniform and macro-uniform color spaces has been coined to make a distinction between these cases [14]. The threshold, or micro-uniform, spaces are clearly the best design for a system's baseline quantization, as evidenced by the DICOM medical imaging GSDF (gray-scale display function), and of which the EOTF (electro-optic transfer function) of SMPTE 2084 (PQ) follows a similar approach. IC_TC_P expands the threshold strategy to include color, and so is also a micro-uniform approach. The CIELAB color space was primarily based around perceptual uniform spacing of Munsell test patches, all being above threshold (9 luminance steps from white to black and typically less than 9 steps from neutral to the maximum saturation) and consequently under-predicts threshold visibility [15]. So, it is a good candidate to describe as a macro-uniform color space. There is current debate on whether the micro-uniform or macro-uniform color space will better predict the kinds of color distortions in complex imagery and of practical interest to business. One goal of this paper is to see how effective the appearance and threshold-based approaches are in predicting the kinds of distortions in the databases. To quantify the performance of the different color difference metrics, we use four standard performance evaluation procedures – Root Mean Square Error, Pearson Linear Correlation Coefficient, Spearman Rank-Order Correlation Coefficient and Outlier Ratio.

Typical color difference metrics are pixel-based operations and results are shown on test patches. In our previous work [5], we measured the performance of various color difference metrics on a database of natural HDR/WCG images. However, applying these metrics on a pixel-by-pixel basis, as done with ΔE_Z and ΔE_{ITP}, without accounting for spatial information from neighboring pixels does not really mimic the Human Visual System (HVS). With such approaches, for the same magnitude of color distortion, a single pixel can have as large of an effect as a large image region if a maximum error criteria is used. Averaging the results across an entire image can be used to avoid this kind of mis-assessment, but may hide large errors that span only a few pixels. Consider two cases of distortion. One is a 100x100 contiguous pixel region while the other has the same number of pixels (10,000) and magnitude but the pixels are scattered across the image individually or in very small regions. The chief difference between these cases, which we expect to result in a different magnitude of visibility, is the spatial frequency. While the spatial blur caused by the optics in the eye has been incorporated into the CSF of some metrics, it is not a complete model of the spatial response of the eye. For example, optical blur is wavelength dependent, as caused by chromatic aberration, and there are neuronal effects on spatio-chromatic visibility that are found in the psycho-physical measurements of spatial frequency sensitivity (CSF) when tested for opponent color signals, which are iso-luminant traverses across the color space. These chromatic CSFs differ from the achromatic (luminance) in that they have lower spatial frequency bandwidth and also are much less band pass than the achromatic CSFs (i.e., more sensitive at lower spatial frequencies [16]). In addition, the CSF for a blue-yellow modulation has less bandwidth than that for the red-green modulations. Lastly, there are complex masking effects across the different achromatic, red-green, and blue-yellow mechanisms [17]. Accordingly, we would like to modify the color representation viz., IC_TC_P and compute the color difference metrics so that it mimics more human visual system behavior to get an improved assessment of HDR image quality.

To account for the limitations of calculating per-pixel color differences, a spatial extension of CIELAB was proposed in S-CIELAB [18], which improved the performance when used with traditional color difference metrics such as ΔE_{00}. In this paper, we propose a similar spatial extension, simulating the spatial filtering by the human visual system (HVS), to the IC_TC_P color representation (which was designed for HDR/WCG content). This approach enables the effects of chromatic aberration on the HVS optics, as well as the neural color opponent differences in behavior (e.g., bandwidths and non-bandpass behavior). However, it does not include the spatio-chromatic effects of masking, which require an substantially increased level of complexity to model.

Methodology

We evaluated three color difference metrics, using the umbrella term ΔE. The ΔE_{00} (CIEDE2000) [19]) is widely used in the industry today, the other two are more recently developed color difference metrics intended to improve performance for HDR/WCG imagery (ΔE_{ITP} [20] and ΔE_Z [21]). For each image / score pair, we calculated the ΔE value between every distorted and reference pixel. Then we averaged over the entire image as shown in Equation 1.

$$\overline{\Delta E_{metric}} = \frac{1}{I*J}\sum_{i=1}^{I}\sum_{j=1}^{J}\Delta E_{metric}(i,j), \tag{1}$$

where (i,j) is the pixel location and $I*J$ are the total number of pixels in the image.

We correlated this average ΔE value (from each image and distortion parameter pair) with the subjective scores of the corresponding images, where better correlation indicates the metric is better at predicting visual quality. We compared other methods to consolidate the set of pixel color differences into a single value (maximum, median) and found that the average results in the best correlation. Brief descriptions of the color difference metrics being evaluated are given below.

Color Difference Metrics

ΔE_{00}

The ΔE_{00} color difference formula is based on the CIE L*a*b* color space. However, this color difference formula was a major revision which included new warp/rotation terms to further improve uniformity of the CIE L*a*b* color space. The majority of the data used to develop this formula was based on paint chips. Despite being developed with reflective data, it is commonly used in SDR emissive display calibration.

ΔE_{00} requires an adapting white point (since it uses CIE L*a*b*). The adapting white point luminance (D65 is used as the chromaticity) we used for ΔE_{00} is 100 cd/m^2. Using an adaptive white point luminance of 1000 cd/m^2 is not ideal [22] and [23] has shown that using a value closer to diffuse white (as opposed to the highlight maximum) produces better results. SDR video commonly placed the reference white point at 100 cd/m^2, and some implementations of HDR place the max diffuse white point in a similar range place as SDR (e.g., 100-200) and save the increased upper range for the specular highlights which can substantially exceed the max diffuse white [24]. Choudhury et al. [5] also showed better performance using a value of 100 cd/m^2. We also compute ΔE_{00} in the S-CIELAB [18] color space, which is the spatial extension of CIELAB color space. Hereafter, we refer to this metric as ΔE_{00}^{S}.

ΔE_Z

ΔE_Z is built upon the $J_z a_z b_z$ color space [21]. The $J_z a_z b_z$ color space was designed for large and small perceptual uniformity (i.e., macro- and micro-uniformity). It utilizes the PQ (ST 2084) transfer function for modeling the achromatic non-linearity of the visual system to improve the HDR performance, but the $J_z a_z b_z$ color space converts from absolute light levels to relative levels and thus normalizes to a concept of "white", with the aim of improving lightness correlation. The lightness correlation optimization was based on data [23] with a diffuse white of 997 cd/m^2. The other parameters were optimized for hue uniformity and perceptual color difference. The ΔE_Z value is calculated through chroma and hue calculations in the polar $J_z a_z b_z$ space as follows -

$$C_z = \sqrt{a_z^2 + b_z^2}, \tag{2}$$

$$H_z = tan^{-1}(\frac{b_z}{a_z}), \tag{3}$$

where a_z and b_z are the color channels in $J_z a_z b_z$ color space, and

$$\Delta E_Z = \sqrt{\Delta J_z^2 + \Delta C_z^2 + \Delta H_z^2} \tag{4}$$

where ΔJ_z, ΔC_z and ΔH_z are the differences between the reference and the distorted image for the J_z, C_z (derived using Equation 2) and H_z (derived using Equation 3) channels respectively. $J_z a_z b_z$ was fit to a number of color patch datsets such as Combined Visual Data (COMBVD) (which in turn includes 4 different datasets - RIT-DuPont, Witt, Leeds and BFD), COMBVD ellipses and Hung and Berns dataset.

ΔE_{ITP}

ΔE_{ITP} is an absolute color difference metric. The adapting white point is not an input into the metric because it has a built in "worst case" adaptation assumption, i.e., best-case visual system performance. It models the case where the viewer is optimally adapted to the image region being evaluated [9]. This metric is standardized in ITU-R BT.2124 [20], is based on the IC_TC_P color representation (intended as an encoding representation), and has been shown to work well for predicting HDR color differences using test patches [10] under rigorous 4AFC threshold test conditions [25, 26]. The IC_TC_P color space also utilizes the PQ (ST 2084) transfer function, applied to the LMS cone signal, motivated by the finding that a cone non-linearity model can predict the PQ (ST 2084)4 non-linearity, when used in a floating adaptation manner [27]. It can be converted into the perceptually uniform ITP color space by following equations -

$$T = C_T * 0.5, \tag{5}$$

$$P = C_P, \tag{6}$$

where the T channel (Tritanopic or blue-yellow) of ITP is obtained by scaling the C_T channel of IC_TC_P and the P channel (Protanopic, or red-green) of ITP color space is the same as the C_P channel of IC_TC_P. The ITP space was optimized to improve hue linearity and small perceptual uniformity.

The ΔE_{ITP} value is calculated in the ITP color space as shown in Equation 7.

$$\Delta E_{ITP} = 720 * \sqrt{\Delta I^2 + \Delta T^2 + \Delta P^2} \tag{7}$$

where ΔI, ΔT and ΔP are the differences between the reference and the distorted image for the I, T and P channels respectively. The ΔE_{ITP} value is calculated from IC_TC_P through scaled Euclidean distance (the 720 scalar and the weights on C_T) to transition from an encoding-based system and to have a value of 1 correlate with visual threshold.

Proposed Color Difference Metric

In this section, we describe the details of the proposed spatial extension of ITP. A high-level overview of the proposed metric is shown in Figure 1. In order to mimic the spatial blurring applied by the human visual system [28], we apply spatial filtering to the color image. Both the reference and the distorted images are first transformed from its native color representation to the ITP color representation as defined in ITU-R BT.2124 [20]. Please note that ITP is an opponent-color space.

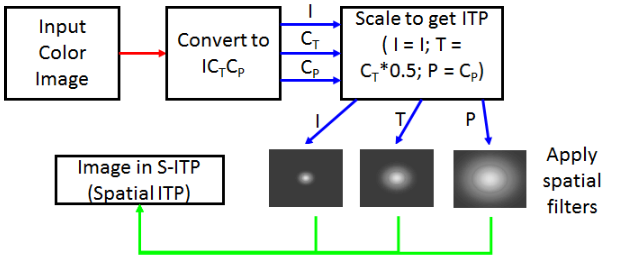

Input Color Image
Convert to IC_TC_P
I
C_T
C_P
Scale to get ITP (I = I; T = C_T*0.5; P = C_P)
I
T
P
Apply spatial filters
Image in S-ITP (Spatial ITP)

Figure 1. Block diagram of proposed spatial extension of ITP. (Please note that the spatial filters are for illustrative purposes and do not necessarily correspond to the actual weights).

Each of the three opponent color channels is convolved with a separable spatial kernel to model the basic effects of the spatial frequency-dependent CSF (contrast sensitivity function). In order to better understand the roles of spatial filtering versus the underlying color space, we use the same spatial components as used for S-CIELAB [18]. These CSF's are simple, being solely a sum of isotropic low pass Gaussian filters specified in the spatial domain as psfs. This can be denoted by Equation 8.

$$f = k\sum_i w_i E_i, \tag{8}$$

where,

$$E_i = z_i e^{-\frac{x^2+y^2}{\sigma_i^2}}. \tag{9}$$

z_i is the scale factor so that E_i sums up to 1. k is the normalization factor so that the two-dimensional kernel f sums up to 1.

The values of w and σ for both the achromatic and the chromatic color channels are as shown in Table 1.

Table 1: Kernel parameters for filtering. w_i are the weights and σ_i is the spread measured in degrees of visual angle

Color Channel	w_i	σ_i
I	0.921	0.0283
	0.105	0.133
	-0.108	4.336
T	0.488	0.0536
	0.371	0.386
P	0.531	0.0392
	0.330	0.494

This spatial convolution is applied to both the reference image as well as the distorted images in the ITP color space. Finally, the spatial version of ΔE_{ITP} (referred to as ΔE^S_{ITP}) is computed as shown in Equation 7, where instead of using the I, T and P color channels of the reference and the distorted images, we use the spatially filtered I, T and P color channels. Note that ΔE^S_{ITP} measures both spatial and color sensitivity.

A brief overview of the method is shown in Algorithm 1.

Algorithm 1 Computing ΔE_{ITP} in spatial extension of ITP color space (ΔE^S_{ITP})

```
Input: $Im_{ref}, Im_{dis}$                          ▷ Reference and Distorted Images
Output: $\Delta E^S_{ITP}$                           ▷ Proposed Color Difference Metric
procedure $\Delta E^S_{ITP}$
    $[I_{ref}, T_{ref}, P_{ref}] = ITP(Im_{ref})$    ▷ Convert to opponent color-space
    $[I_{dis}, T_{dis}, P_{dis}] = ITP(Im_{dis})$    ▷ Convert to opponent color-space
    Filter $I_{ref}$, $T_{ref}$ and $P_{ref}$ opponent color channels of $Im_{ref}$ using Equations 8 and 9 and parameters from Table 1    ▷ Spatial filtering of channels
    Filter $I_{dis}$, $T_{dis}$ and $P_{dis}$ opponent color channels of $Im_{dis}$ using Equations 8 and 9 and parameters from Table 1    ▷ Spatial filtering of channels
    Calculate $\Delta E^S_{ITP}(x,y)$ as shown in Equation 7 where $\Delta I(x,y) = I_{ref}(x,y) - I_{dis}(x,y)$, $\Delta T(x,y) = T_{ref}(x,y) - T_{dis}(x,y)$ and $\Delta P(x,y) = P_{ref}(x,y) - P_{dis}(x,y)$ and the pixel location is $(x,y)$
    Calculate $\Delta E^S_{ITP}$ as shown in Equation 1    ▷ Mean value across entire image
end procedure

function ITP($Im$)                                   ▷ Calculate ITP color channels
    Convert display-referred linear R, G, B (in accordance with Table 10 of Recommendation ITU-R BT. 2100 [29]) to linear L, M, S (in accordance with Table 7 of Recommendation ITU-R BT. 2100 [29])
    Convert linear L, M, S to non-linear L', M', S' by applying the PQ non-linearity defined in Table 4 of Recommendation ITU-R BT. 2100 [29]
    Convert non-linear L', M', S' to I, $C_T$, $C_P$ as defined in Table 7 of Recommendation ITU-R BT. 2100 [29]
    Scale I, $C_T$, $C_P$ channels to create $I$, $T$, $P$ channels as shown in Equations 5 and 6    ▷ I channel stays the same
    return $[I, T, P]$
end function
```

Databases

We considered four publicly available databases from different labs comprised entirely of natural images to compare the performance of the different color difference metrics for evaluation. The digital images and subjective scores are made available for independent researchers to do various analysis and metric development. The first database [30] (Database 1) contains 20 reference HDR images. Distorted images were created by compressing the reference images using JPEG XT with various profiles and quality levels. Two different tone mapping operations [31, 32] were used for the base layer. Four different bit rates were chosen using three profiles of JPEG XT. Each image had a resolution of 944

X 1080 pixels (i.e., a crop for a split-screen of a 1920x1080) and were calibrated for a SIM2 HDR monitor.

The second database [33] that we considered is a combination of two different databases [33, 34]. One of them [34] is composed of five original HDR images which were first tone-mapped [35], from which 50 compressed images were obtained using three different coding schemes - JPEG, JPEG2000 and JPEG XT. These images were presented one after the other (sequentially) on a SIM2 HDR47E display and scores were collected from 15 participants. The second database [33] also uses five original HDR images from which 50 compressed images were obtained, using JPEG and JPEG2000 (with different bit rates), as well as additional SDR images obtained using two different mapping operations [35, 36]. The second database (Database 2) has 100 1920 X 1080 images.

One of the limitations of the two databases mentioned above was that these databases did not explore wider color gamut. Also, they did not specifically contain any color artifacts, although some may arise from tone-mapping at the very top and bottom of the color solid (and if any chromatic sub-sampling was used in the compression profiles). In addition, the subjective testing for all three above-mentioned databases were conducted on the same monitor (SIM2 HDR monitor). To introduce more variety in our experimental samples we added two additional databases (Database 3 and 4) that included chromatic distortions as well as images that extended beyond the ITU-R BT.709 color gamut (up to ITU-R BT.2020 color gamut by use of the Sony BVM-X300 OLED professional monitor).

The third database [37] (Database 3) contains eight images distorted using four types of distortion: HEVC compression using four different quantization parameters (QP), HEVC compression without the chroma QP adaptation resulting in chromatic distortions at three different values, three different levels of Gaussian noise and two different types of gamut mismatch (i.e., rendered assuming that ITU-R BT.709 images were interpreted as ITU-R BT.2020 images leading to more saturated colors, and assuming that ITU-R BT.2020 images were interpreted as ITU-R BT.709 images leading to less saturated colors).

The fourth database [38] (Database 4) contains eight images that were compressed with four different QP using three different compression options – Recommended HEVC compression, HEVC compression without chroma QP offset algorithm and HEVC compression with 8 bits quantization for chroma instead of 10 during compression. These images are all represented using the ITU-R BT.2020 color gamut.

Experimental Results and Discussion

In this section, we compare the relative performance of various color difference metrics and their proposed spatial extensions on the four databases mentioned in the previous section. We evaluated the performance of the metrics by comparing the subjective scores with the scores predicted from the different metrics using a standardized method [39] used by the video quality experts group (VQEG). In that standard approach, a monotonic logistic function is used to fit the objective prediction to the subjective scores as follows:

$$f = \alpha + \frac{\beta}{1 + e^{-\gamma\cdot(x-\delta)}}, \tag{10}$$

Table 2: Performance comparison on Database 1 [30]

Method	PLCC	SROCC	RMSE	OR
ΔE_{00}	0.7946	0.7901	0.7644	0.6458
ΔE_Z	0.6672	0.6717	0.9383	0.7375
ΔE_{ITP}	0.8366	0.8379	0.6878	0.6375
ΔE^S_{00}	0.8773	0.8760	0.6030	**0.5708**
$\boldsymbol{\Delta E^S_{ITP}}$	**0.8995**	**0.8980**	**0.5479**	0.5917

Table 3: Performance comparison on Database 2 [33, 34]

Method	PLCC	SROCC	RMSE	OR
ΔE_{00}	0.6134	0.5991	23.5999	0.75
ΔE_Z	0.5382	0.5145	25.1613	0.78
ΔE_{ITP}	0.7148	0.7290	20.8168	0.74
ΔE^S_{00}	0.7209	0.7433	20.9077	0.68
$\boldsymbol{\Delta E^S_{ITP}}$	**0.8224**	**0.8183**	**17.1490**	**0.63**

where f is the fitted objective score, x is the predicted score using different techniques and $\alpha, \beta, \gamma, \delta$ are the parameters that define the shape of the logistic fitting function.The fit is computed by minimizing the least squares error between the subjective and the fitted objective scores. This mapping function is used to mimic the fact that high-level cognitive processes are required to map the lower-level perceptions to a score. The rationale is that the various metrics can model low-level perception, but that high-level cognitive processes are required to arrive at a score. As a simplified model of this internal mapping step, the logistic function with variable parameters is currently being used as a surrogate until better understanding is achieved. Please note that the subjective scores for each database have been made available by the respective authors.

We use the following four standard evaluation procedures and criteria [39] to measure the performance – Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) for measuring prediction accuracy, Spearman Rank-Order Correlation Coefficient (SROCC) for prediction monotonicity and Outlier Ratio (OR) to determine prediction consistency. Lower values of RMSE and OR, and higher values of PLCC and SROCC indicates better performance.

We report the performance of the different color difference metrics in Tables 2, 3, 4 and 5. The best metric for each database along with the best scores are highlighted in bold. We refer to the color difference metrics based on pixel-wise calculations as ΔE_{00}, ΔE_Z and ΔE_{ITP}. For each metric, we compute the average ΔE value as shown in Equation 1 and compare that with the subjective scores of the corresponding images using each of the performance indices (PLCC, SROCC, RMSE and OR) mentioned in the previous section. Please refer to our previous work [5], where we have conducted an in-depth analysis of the comparison of the pixel-wise color difference metrics viz., ΔE_{00}, ΔE_Z and ΔE_{ITP}. In this paper, we will focus more on the performance of the spatial extension of the color difference metrics. Specifically, we compare the proposed color difference metric (hereafter referred to as ΔE^S_{ITP}, which is the spatial extension of ΔE_{ITP}) with the spatial version of CIELAB (S-CIELAB) using which we computed the ΔE_{00} metric (referred to as ΔE^S_{00}).

Table 4: Performance comparison on Database 3 [37]

Method	PLCC	SROCC	RMSE	OR
ΔE_{00}	0.2738	0.2191	22.7381	0.6042
ΔE_{Z}	0.3046	0.2966	22.5173	0.6667
ΔE_{ITP}	0.3901	**0.3208**	21.7588	0.6563
ΔE^{S}_{00}	0.3873	0.2244	21.7887	0.6354
$\boldsymbol{\Delta E^{S}_{ITP}}$	**0.4787**	0.2764	**20.7473**	**0.5938**

Table 5: Performance comparison on Database 4 [38]

Method	PLCC	SROCC	RMSE	OR
ΔE_{00}	0.3983	0.3119	20.3594	0.5938
ΔE_{Z}	0.3294	0.2880	20.9594	0.7083
ΔE_{ITP}	0.6932	0.6982	16.0149	0.5833
ΔE^{S}_{00}	0.6307	0.6174	17.2259	0.6354
$\boldsymbol{\Delta E^{S}_{ITP}}$	**0.8710**	**0.8643**	**10.9033**	**0.3646**

We can see that overall the best performance is obtained using the proposed metric, ΔE^{S}_{ITP}. Also, both of the spatial versions of the color difference metrics consistently improved the performance over their corresponding pixel-wise color difference metrics. This implies that spatial filtering of the achromatic and chromatic channels improves the prediction. A few variations of spatial filtering were tested with the ITP color space: i) Filtering only the I channel before computing ΔE_{ITP} (ΔE^{S}_{ITP}_I), and ii) Filternig the per-pixel results after computing ΔE_{ITP} (ΔE^{S}_{ITP}_*Post*). The results are shown in Table 6. Neither of these variations improves upon the per-pixel results (Comparing Tables 2 and 6). This implies that spatial filtering of individual chromatic channels before computing the color difference metric is an important component of the metric. Note that neither of these two variations are physiologically correct in terms of where the spatial low-pass filtering of achromatic and chromatic operations are occurring in the neural pathway. Thus, we have further support for the overall physiological models that the ΔE^{S}_{ITP} and ΔE^{S}_{00} metrics are based on.

Table 6: Variation of spatial filtering on ΔE_{ITP} on Database 1

Method	PLCC	SROCC	RMSE	OR
ΔE^{S}_{ITP}_I	0.8367	0.8378	0.6876	0.6375
ΔE^{S}_{ITP}_*Post*	0.8365	0.8376	0.6879	0.6375

We observe that the overall performance of all the metrics is poorer on Database 3 compared to other databases. This might be due to the fact that Database 3 has a wide variety of artifacts. Some distortions such as gamut mismatch might be clearly visible but not associated with loss of quality for some viewers. These kinds of color distortions may look plausible to the viewer even if incorrect and thus may not be penalized as much as other distortions. In this paper, we used the same filtering parameters as S-CIELAB for the spatial extension of ITP to compute ΔE^{S}_{ITP}. Optimizing these parameters might result in improved performance. On the other hand, Database 1 seems less selective and most metrics already have very high correlation and low error on that database. Using the spatial version results in relatively less improvement on Database 1 as compared to the other databases.

Table 6: Statistical significance on relative performance between PLCC values of the metrics on Database 1

Method	ΔE_{00}	ΔE_{Z}	ΔE_{ITP}	ΔE^{S}_{00}	ΔE^{S}_{ITP}
ΔE_{00}	-	1	0	-1	-1
ΔE_{Z}	-1	-	-1	-1	-1
ΔE_{ITP}	0	1	-	0	-1
ΔE^{S}_{00}	1	1	0	-	0
ΔE^{S}_{ITP}	1	1	1	0	-

On databases 1 and 2 which predominantly contain compression artifacts, we found that although ΔE_{00} had worse prediction than ΔE_{ITP}, using ΔE^{S}_{00} results in improved performance over using ΔE_{ITP}. That is, the spatial filtering had a stronger impact than the color space. On the other hand, for databases 3 and 4 that specifically contain chromatic artifacts, we found that using just ΔE_{ITP} already outperforms ΔE^{S}_{00}. In this case, the color space had more impact than the filtering. However, overall using ΔE^{S}_{ITP} results in the best performance. We have already shown the advantages of using ITP over CIELAB [5]. Likewise, we can see benefits of using the spatial filtering on HDR-WCG color space of ITP over S-CIELAB by using a similar spatial filtering.

Rousselot et al. [38] also find similar trends in performance with regards to color difference metrics (they found a precursor to ΔE_{ITP} outperforming ΔE_{00} and ΔE_{Z}). That version didn't achieve as high correlation values (at least with regards to ΔE_{ITP}) as our approach. For instance, they report a PLCC of 0.8065 on Database 1 compared to our PLCC of 0.836 using DEITP. It is unclear if the slight difference in the metric caused the difference, or other possible unstated assumptions in their calculations. Rousselot et al. [38] also reported results using sophisticated HDR metrics (HDR-VDP-2 [40, 41] and HDR-VQM [42]) on the 4 databases. We observe that for database 4 [38] that primarily contains chromatic distortions, using the proposed color difference metric, ΔE^{S}_{ITP} outperforms both HDR-VDP-2 (PLCC = 0.8605, RMSE = 11.3) and HDR-VQM (PLCC = 0.7714, RMSE = 14.11). However, on the other databases, both HDR-VDP-2 and HDR-VQM still outperform ΔE^{S}_{ITP}.

Statistical Analysis

To evaluate whether the difference between the performance of two different color difference metrics is statistically significant, we performed statistical tests (Z-test using Fisher z-transformation) according to the recommendations proposed in ITU-T P.1401 [43]. The statistical significance between the color difference metrics for Databases 2 through 5 in listed in Tables 7 through 10 respectively. In these tables, the symbols "1", "0" or "-1" respectively indicate that the corresponding row metric is statistically (with 95% confidence) superior, equivalent or inferior than the column metric. Please note that the matrices that are shown in Tables 7 through 10 are all skew-symmetric matrices and the lower triangular and the upper triangular matrices have the same interpretation. We obtained similar trends for the other evaluation criteria – SROCC, RMSE and OR, and do not report those results in this paper.

We observe that on Database 1, ΔE^{S}_{00} is statistically superior to both ΔE_{00} and ΔE_{Z} but is equivalent to ΔE_{ITP}. On the other hand, ΔE^{S}_{ITP} is superior to all pixel-wise color difference metrics

Table 7: Statistical significance on relative performance between PLCC values of the metrics on Database 2

Method	ΔE_{00}	ΔE_Z	ΔE_{ITP}	ΔE^S_{00}	ΔE^S_{ITP}
ΔE_{00}	-	0	0	0	-1
ΔE_Z	0	-	-1	-1	-1
ΔE_{ITP}	0	1	-	0	-1
ΔE^S_{00}	0	1	0	-	-1
ΔE^S_{ITP}	1	1	1	1	-

Table 8: Statistical significance on relative performance between PLCC values of the metrics on Database 3

Method	ΔE_{00}	ΔE_Z	ΔE_{ITP}	ΔE^S_{00}	ΔE^S_{ITP}
ΔE_{00}	-	0	0	0	-1
ΔE_Z	0	-	0	0	0
ΔE_{ITP}	0	0	-	0	0
ΔE^S_{00}	0	0	0	-	0
ΔE^S_{ITP}	1	0	0	0	-

Table 9: Statistical significance on relative performance between PLCC values of the metrics on Database 4

Method	ΔE_{00}	ΔE_Z	ΔE_{ITP}	ΔE^S_{00}	ΔE^S_{ITP}
ΔE_{00}	-	0	-1	-1	-1
ΔE_Z	0	-	-1	-1	-1
ΔE_{ITP}	1	1	-	0	-1
ΔE^S_{00}	1	1	0	-	-1
ΔE^S_{ITP}	1	1	1	1	-

but it is equivalent to ΔE^S_{00}. On database 2, amongst the pixel-wise metrics, ΔE_{ITP} is equivalent to ΔE_{00} but superior to ΔE_Z. ΔE^S_{00} is statistically equivalent to both ΔE_{00} and ΔE_{ITP}. However, ΔE^S_{ITP} is statistically superior to all other color difference metrics. Although ΔE^S_{ITP} outperforms all the color difference metrics on database 3 (Table 4), its performance is statistically equivalent to the other color difference metrics with the exception of ΔE_{00}, where its performance is superior. One of the reasons that the performance of ΔE^S_{ITP} is not superior on Database 3 could be because the size of database is relatively small (96 images). In addition, Database 3 was the one where all the metrics performed least well, possibly as a result of the widely varying color distortions requiring an even more advanced model than simple spatial filtering of color channels. On database 4 we found that ΔE_{ITP} is superior to the other two pixel-wise color difference metrics. However, it is equivalent to ΔE^S_{00} although the performance is better. Statistically, ΔE^S_{ITP} is superior to all other color difference metrics. This statistical analysis shows the actual improvement using the proposed metric.

Conclusion & Future Work

In this paper we report on the performance of several color difference metrics and their spatial extensions to assess the quality of a variety of HDR image distortions. Specifically, we compare ΔE_{00} with two other color difference metrics designed for HDR/WCG images: ΔE_Z and ΔE_{ITP}. We compute ΔE^S_{00} in the S-CIELAB color space, which is the spatial extension of the CIELAB color space. We also propose our new metric, ΔE^S_{ITP}, which is a similar extension to apply spatial filtering to the ITP color space. We evaluated the metrics using four different databases containing a wide variety of distortions and show that overall, ΔE_{ITP} outperforms the other pixel-wise color difference metrics on all four databases. For example, ΔE_{ITP} has 73% improvement over ΔE_{00} for database 4, and for database 1, which is the the database of most similar performance, it has an improvement of 5% over ΔE_{00}. We also show improvement in performance while using ΔE_{00} in the S-CIELAB color space over the CIELAB color space, although it does not always outperform ΔE_{ITP}. Finally we show that our proposed metric ΔE^S_{ITP}, which works in the spatial extension of ITP color representation results in a marked improvement in prediction over all other metrics tested. Finally, we performed a statistical analysis to validate the effectiveness of ΔE^S_{ITP} being used in the measurement of perceptual image quality. In the Introduction, the micro-uniform and macro-uniform types of color spaces was described, with the question of which gives better results for distortions in complex imagery (e.g., of business interest). That complex images are generally supra-threshold suggests a macro-uniform color space would be more relevant for such imagery. However, these results show that the micro-uniform color space concept gives better predictions of the quality ratings of complex imagery (at least for these databases). It is possible that while the main content of the imagery is well into the ranges of supra-threshold imagery, the differences between the images, that is, the aspects that leads to the subjective ratings are better described by threshold models.

Future work suggested by these results indicate that a more advanced spatio-chromatic modeling and filtering could further improve performance. In addition, generating additional image quality databases with ever improving display technologies, improved psycho-physical design, and more statistically characterized imagery will help close the circle on the development of highly useful spatio-chromatic HDR-WCG quality metrics.

Acknowledgments

We would like to thank Robin Atkins for his valuable suggestions.

References

[1] Lahoulou, A., Larabi, M. C., Beghdadi, A., Viennet, E., and Bouridane, A., "Knowledge-based taxonomic scheme for full-reference objective image quality measurement models," *Journal of Imaging Science and Technology* **60**(6), 60406–1–60406–15 (2016).

[2] Keelan, B. W., "Predicting multivariate image quality from individual perceptual attributes," in [*PICS 2002: IS&T's PICS Conference, An International Technical Conference on Digital Image Capture and Associated System, Reproduction and Image Quality Technologies, April 2002, Portland, Oregon*], 82–87 (2002).

[3] Choi, S., Luo, M., Pointer, M., and Rhodes, P., "Investigation of large display color image appearance i: Important factors affecting perceived quality," *Journal of Imaging Science and Technology - J IMAGING SCI TECHNOL* **52** (07 2008).

[4] Dror, R. O., Willsky, A. S., and Adelson, E. H., "Statisti-

cal characterization of real-world illumination," *Journal of Vision* **4**, 11–11 (09 2004).

[5] Choudhury, A., Pytlarz, J., and Daly, S., "HDR and WCG image quality assessment using color difference metrics," in [*SMPTE 2019 Annual Technical Conference and Exhibition*], (Oct 2019).

[6] Miller, S., Nezamabadi, M., and Daly, S., "Perceptual signal coding for more efficient usage of bit codes," *SMPTE Motion Imaging Journal* **122**, 52–59 (May 2013).

[7] Aydın, T. O., Mantiuk, R., and Seidel, H.-P., "Extending quality metrics to full dynamic range images," in [*Human Vision and Electronic Imaging XIII*], *Proceedings of SPIE*, 6806–10 (January 2008).

[8] Susstrunk, S. and Finlayson, G. D., "Evaluating chromatic adaptation transform performance," *Proc. IS&T/SID 13th Color Imaging Conference* , 75–78 (2005).

[9] Kunkel, T., Wanat, R., Pytlarz, J., Pieri, E., Atkins, R., and Daly, S., "Assessing color discernibility in hdr imaging using adaptation hulls," *Color and Imaging Conference* **2018**(1), 336–343 (2018).

[10] Pieri, E. and Pytlarz, J., "Hitting the mark - a new color difference metric for hdr and wcg imagery," in [*SMPTE 2017 Annual Technical Conference and Exhibition*], 1–13 (Oct 2017).

[11] Hillis, J. M. and Brainard, D. H., "Do common mechanisms of adaptation mediate color discrimination and appearance? uniform backgrounds," *J. Opt. Soc. Am. A* **22**, 2090–2106 (Oct 2005).

[12] Hillis, J. M. and Brainard, D. H., "Do common mechanisms of adaptation mediate color discrimination and appearance? contrast adaptation," *J. Opt. Soc. Am. A* **24**, 2122–2133 (Aug 2007).

[13] Lissner, I. and Urban, P., "How perceptually uniform can a hue linear color space be?," in [*Color Imaging Conference*], (2010).

[14] Holm, J., "Some considerations in quantifying display color volume," in [*QLED and HDR10+ Summit*], (2017).

[15] Meninger, C. and Pruitt, T., "Deitp is now itu-r bt.2124–is the industry ready to move on from de2000," in [*SMPTE 2019 Annual Technical Conference and Exhibition*], (Oct 2019).

[16] Mullen, K. T., "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings.," *The Journal of Physiology* **359**(1), 381–400 (1985).

[17] Switkes, E., Bradley, A., and Valois, K. K. D., "Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings," *J. Opt. Soc. Am. A* **5**, 1149–1162 (Jul 1988).

[18] Zhang, X. and Wandell, B. A., "A spatial extension of cielab for digital color-image reproduction," *Journal of the Society for Information Display* **5**(1), 61–63 (1997).

[19] ISO/CIE 11664-6:2014(E), "Colorimetry - Part 6: CIEDE2000 Colour-Difference Formula," Standard (2014).

[20] ITU-R BT. 2124, "Objective metric for the assessment of the potential visibility of colour differences in television," Standard (Jan 2019).

[21] Safdar, M., Cui, G., Kim, Y. J., and Luo, M. R., "Perceptually uniform color space for image signals including high dynamic range and wide gamut," *Opt. Express* **25**, 15131–15151 (Jun 2017).

[22] Reinhard, E., Stauder, J., and Kerdranvat, M., "An assessment of reference levels in hdr content," *SMPTE Motion Imaging Journal* **128**, 20–27 (April 2019).

[23] Fairchild, M. D. and Chen, P.-H., "Brightness, lightness, and specifying color in high-dynamic-range scenes and images," in [*Image Quality and System Performance VIII*], Farnand, S. P. and Gaykema, F., eds., **7867**, 233 – 246, International Society for Optics and Photonics, SPIE (2011).

[24] Wolff, "On the relative brightness of specular and diffuse reflection," in [*1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*], 369–376 (June 1994).

[25] Jakel, F. and Wichmann, F. A., "Spatial four-alternative forced-choice method is the preferred psychophysical method for naive observers," *Journal of Vision* **6**, 13–13 (11 2006).

[26] Krauskopf, J. and Gegenfurtner, K., "Color discrimination and adaptation," *Vision Research* **32**(11), 2165 – 2175 (1992).

[27] Daly, S. and Golestaneh, S. A., "Use of a local cone model to predict essential CSF light adaptation behavior used in the design of luminance quantization nonlinearities," in [*Human Vision and Electronic Imaging XX*], Rogowitz, B. E., Pappas, T. N., and de Ridder, H., eds., **9394**, 16 – 26, International Society for Optics and Photonics, SPIE (2015).

[28] Johnson, G. M. and Fairchild, M. D., "On contrast sensitivity in an image difference model," in [*PICS 2002: IS&T's PICS Conference*], 18–23 (2002).

[29] ITU-R BT. 2100-2, "Image parameter values for high dynamic range television for use in production and international programme exchange," Standard (July 2018).

[30] Korshunov, P., Hanhart, P., Richter, T., Artusi, A., Mantiuk, R., and Ebrahimi, T., "Subjective quality assessment database of HDR images compressed with jpeg xt," in [*QoMEX*], 1–6 (May 2015).

[31] Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "A perceptual framework for contrast processing of high dynamic range images," *ACM Trans. Appl. Percept.* **3**, 286–308 (July 2006).

[32] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," *ACM Trans. Graph.* **21**, 267–276 (July 2002).

[33] Zerman, E., Valenzise, G., and Dufaux, F., "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience* **2**, 5 (Apr 2017).

[34] Valenzise, G., Simone, F. D., Lauga, P., and Dufaux, F., "Performance evaluation of objective quality metrics for hdr image compression," in [*SPIE optical engineering + applications, International Society for Optics and Photonics*], (2014).

[35] Mai, Z., Mansour, H., Mantiuk, R., Nasiopoulos, P., Ward, R., and Heidrich, W., "Optimizing a tone curve for backward-compatible high dynamic range image and video compression," *IEEE Transactions on Image Processing* **20**, 1558–1571 (June 2011).

[36] Miller, S., Nezamabadi, M., and Daly, S., "Perceptual signal coding for more efficient usage of bit codes," in [*The 2012*

Annual Technical Conference Exhibition], 1–9 (Oct 2012).

[37] Rousselot, M., Auffret, E., Ducloux, X., Le Meur, O., and Cozot, R., “Impacts of viewing conditions on hdr-vdp2,” in [*EUSIPCO*], 1442–1446 (Sept 2018).

[38] Rousselot, M., Le Meur, O., Cozot, R., and Ducloux, X., “Quality assessment of hdr/wcg images using hdr uniform color spaces,” *Journal of Imaging* **5**(1) (2019).

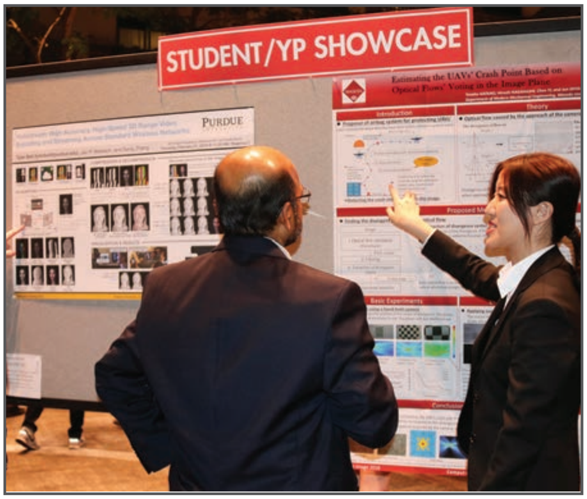
[39] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” (2003).

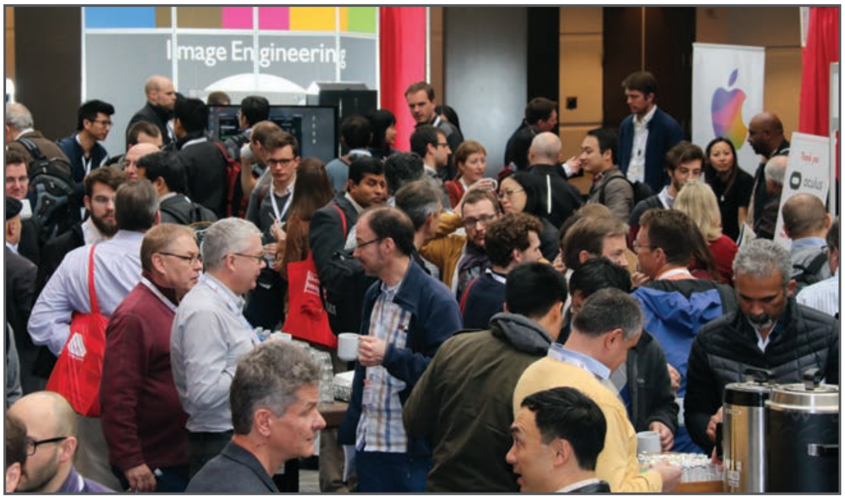
[40] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W., “HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions,” *ACM Trans. Graph.* **30**, 40:1–40:14 (July 2011).

[41] Narwaria, M., Mantiuk, R., Silva, M. P. D., and Callet, P. L., “HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images,” *Journal of Electronic Imaging* **24**, 24 – 24 – 3 (2015).

[42] Narwaria, M., Silva, M. P. D., and Callet, P. L., “HDR-VQM: An Objective Quality Measure for High Dynamic Range Video,” *Signal Processing: Image Communication* **35**, 46–60 (July 2015).

[43] ITU-T, “Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models,” Standard (2012).





- AR is a fundamental and transformative interface and platform.
- 20+ years from my initial working demo to today's AR displays
- Optimistic that it won't take another 20 to reach ubiquitous consumer AR!