Analysis of Public Image and Video Databases for Quality Assessment

Stefan Winkler

***Abstract*—Databases of images or videos annotated with subjective ratings constitute essential ground truth for training, testing, and benchmarking algorithms for objective quality assessment. More than two dozen such databases are now available in the public domain; they are presented and analyzed in this paper. We propose several criteria for quantitative comparisons of source content, test conditions, and subjective ratings, which are used as the basis for the ensuing analyses and discussion. This information will allow researchers to make more well-informed decisions about databases, and may also guide the creation of additional test material and the design of future experiments.**

***Index Terms*—Image and video quality assessment, mean opinion score (MOS), peak signal-to-noise ratio (PSNR), subjective experiments.**

I. Introduction

Ground truth is one of the most important and useful components for the evaluation and benchmarking of new algorithms. In the field of image and video quality, ground truth means databases of test clips annotated with subjective ratings. The lack of annotated databases used to be a major hurdle for researchers working on quality assessment algorithms. Even uncompressed video content was hard to find. In recent years, an increasing a number of databases have been released into the public domain, to the point where it has become be hard to keep track of them or to choose the most suitable one. For this paper, we have compiled a list of 27 image and video databases that are publicly available and relevant to quality assessment.

Comparing databases using the same criteria is helpful for model developers, who can make a more informed decision about which databases may be most suited for their specific benchmarking or other needs. Furthermore, by providing an overview of what is currently available in a uniform framework, this study highlights areas where additional databases are needed and where researchers may want to focus in the design of future experiments.

The paper is organized as follows. Section II provides an overview of available image and video databases annotated with subjective quality ratings as well as several additional databases of relevance for the quality assessment community. Section III proposes various criteria for quantitative analyses of source content, test material, and subjective ratings, which are then used for comparing databases. Section IV reviews the analysis findings and discusses areas for database improvement and future work. Section V concludes the paper.

Manuscript received November 09, 2011; revised May 10, 2012; accepted August 06, 2012. Date of publication August 23, 2012; date of current version September 12, 2012. This work was supported by a research grant for ADSC's Human Sixth Sense Program from Singapore's Agency for Science, Technology and Research (A*STAR). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Amy Reibman.

The author is with the Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign (UIUC), Singapore 639928 (e-mail: stefan.winkler@adsc.com.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2012.2215007

II. Databases

This section presents 27 image and video databases that are annotated with subjective quality ratings. An overview of the test material in each database, including the number of sources and test conditions, resolution and format, can be found in Table I. Experiment details, such as subjective testing methods and data, subjects, viewing setup, are provided in Table II. Note that many databases do not provide all these details, as evidenced by the empty table entries. Additional database-specific information, including references and a short description of test conditions, is given in the following. A selection of other databases that are of interest to the image and video quality research community are also mentioned.

The author maintains an up-to-date list of links to these and other databases on his web site [3].

A. Grayscale Images

- *A57 Database* [4]. Small database with various distortion types (compression, blur, noise).
- *IRCCyN/IVC Watermarking Databases* [5]–[7]. Four separate databases created by embedding watermarks with different algorithms. Includes Broken Arrow (BA), Fourier Subband (FSB), Enrico, and Meerwald (MW) databases.
- *Wireless Imaging Quality (WIQ) Database* [8], [9]. JPEG compressed images and distortions introduced by a simulated wireless link.

B. Color Images

- *Categorical Subjective Image Quality (CSIQ) Database* [10], [11]. Distortions include JPEG and JPEG2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blur.
- *IRCCyN/IVC Image Quality Database* [12], [13]. Test conditions include JPEG, JPEG2000, and LAR (Locally Adaptive Resolution) compression as well as blur.
- *IRCCyN/IVC 3D Image Quality Database* [14], [15] was the first public-domain database on 3D image quality. Test conditions include JPEG and JPEG2000 compression as well as blur.

TABLE I
TEST MATERIAL SUMMARY

Database	Type	Year	Total	Rated	SRC	HRC	Resolution	Framerate	Format
A57	Gray	2007	57	54	3	18	512×512		BMP
BA	Gray	2009	130	120	10	12	512×512		PGM
FSB	Gray	2009	215	210	5	42	512×512		BMP
Enrico	Gray	2007	105	100	5	20	512×512		BMP
MW	Gray	2009	132	120	12	10	512×512		BMP
WIQ	Gray	2009	87	80	7		512×512		BMP
CSIQ	Image	2010	930	866	30	24-30	512×512		PNG
IVC(I)	Image	2005	195	185	10	10-25	512×512		BMP
IVC-3D	Image	2008	96	90	6	16	512×512		BMP
IVC-Art	Image	2009	128	120	8	15	512×512		PPM+JPG
LIVE(I)	Image	2006	1011	779	29	25-27	∼768×512		BMP
MICT	Image	2008	196	196	14	12	768×512		BMP
MMSP-3D(I)	Image	2010	54	54	9	6	1920×1080		JPG
TID	Image	2008	1725	1700	25	68	512×384		BMP
EPFL/PoliMI	Video	2009	156	156	12	12	CIF/4CIF	25/30fps	YUV+264
IVC-1080i	Video	2008	192	192	24	7	1080i	25fps	YUV
IVC-RoI	Video	2009	84	84	6	14	576i	25fps	YUV+264
IVP	Video	2011	138	138	10	10-14	1080p	25fps	YUV
LIVE(V)	Video	2010	160	150	10	15	768×432	25/50fps	YUV+264/M2V
MMSP-3D(V)	Video	2010	60	30	6	5	∼1080p	25fps	AVI(XVID)
MMSP-SVD	Video	2010	58	84	3	various	720p	50fps	SVC+264
NYU-1	Video	2008	75	60	6	5	CIF/QCIF	30fps	YUV
NYU-2	Video	2009	68	68	4	16	CIF/QCIF	30fps	YUV
NYU-3	Video	2010	210	180	6	15	CIF/QCIF	30fps	YUV
NYU-PL	Video	2007	34	12	17	1	QVGA	10-15fps	YUV
VQEG-FR	Video	2000	360	320	20	16	480i/576i	25/30fps	UYVY
VQEG-HD	Video	2010	740	740	49	75	1080i/p	25/30fps	AVI(UYVY)

Total — Total number of images or videos.
Rated — Number of images or videos with subjective ratings.
SRC — Number of source (reference) images/videos.
HRC — Number of test conditions (a.k.a. hypothetical reference circuits).
Resolution — Image/video resolution (i/p indicates interlaced/progressive)
Format — Image/video file encoding/format.

- *IRCCyN/IVC Art Image Quality Database* [16], [17]. Half of this dataset are "art" images from museums, such as digitized paintings or photos of sculptures. Test conditions include JPEG, JPEG2000, and LAR compression.
- *LIVE Image Quality Assessment Database* [18], [19]. Release 2 of this popular database. Distortions include JPEG, JPEG2000, white noise, Gaussian blur, and simulated Rayleigh fading channel (JPEG2000 bitstream).
- *MICT (a.k.a. Toyama) Image Quality Evaluation Database* [20] focuses on JPEG and JPEG2000 compression. Two subjective datasets are available, from CRT and LCD monitors [21].
- *MMSP 3D Image Quality Assessment Database* [22], [23]. The test conditions represent different inter-camera distances. All images are JPEG-compressed.
- *Tampere Image Database (TID)* [24], [25]. Currently the largest image quality database available in the public domain, both in terms of test images and number of subjects. It contains a wide variety of distortions, including various types of noise, blur, JPEG and JPEG2000 compression, transmission errors, local image distortions, luminance and contrast changes.

C. Video

- *EPFL/PoliMI Video Quality Assessment Database* [26], [27]. Test conditions focus on H.264 compressed videos corrupted by simulated packet loss due to transmission over an error-prone network.
- *IRCCyN/IVC 1080i Database* [28], [29] comprises high-definition (HD) video compressed using H.264. In addition to ACR MOS [2], SAMVIQ MOS [30] is available for part of the database.
- *IRCCyN/IVC SD RoI Database* [31], [32] includes standard-definition (SD) video compressed using H.264, with and without transmission errors.
- *IVP Database* [33] comprises progressive HD video compressed with MPEG-2, Dirac wavelet, and H.264 codecs as well as H.264 streams affected by simulated packet loss. DMOS are provided separately for expert and non-expert observers.
- *LIVE Video Quality Database* [34], [35]. Test conditions include MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP wired and wireless networks.
- *MMSP 3D Video Quality Assessment Database* [36], [37] is the first public-domain database on 3D video quality. The test conditions represent different camera distances. All videos are slightly cropped and compressed.
- *MMSP Scalable Video Database (SVD)* [38]–[40]. The test conditions include two scalable video codecs using multiple spatial and temporal resolutions. The database only includes the sources together with the software and process for creating the test conditions, rather than including the test videos as such. Subjects performed paired comparisons in side-by-side viewing sessions.

TABLE II
SUBJECTIVE EXPERIMENT SUMMARY

Database	Method	Data	Subjects	Ratings	Age	Female	Screen	Distance	PSNR
A57		DMOS							63%
BA	DSIS	Raw	17	17				6 H_s	93%
Enrico	DSIS	Raw	16	16				6 H_s	
FSB	DSIS	Raw	7	7				6 H_s	74%
MW	DSIS	Raw	14	14				6 H_s	
WIQ	DSCQS	Raw	60	30	20-53	20%	17" CRT/LCD	4-6 H_s	64%
CSIQ	Custom	DMOS+σ	25	5-7	21-35		LCD	80 cm	
IVC(I)	DSIS	DMOS+σ	15	15				6 H_s	65%
IVC-3D	SAMVIQ	DMOS	19	19	μ=28		21" CRT	4 H_s	
IVC-Art	DSIS	Raw	19	19			CRT	4 H_p	
LIVE(I)	ACR	DMOS+'σ'		20-29	students		21" CRT	2-2.5 H_s	88%
MICT	ACR	Raw	16	16			17" CRT	4 H_p	61%
MICT/IVC	ACR	MOS+σ	27	27			17" LCD	4 H_p	62%
MMSP-3D(I)	ACR	Raw	17	17	22-53	6%	46" LCD	3 H_s	
TID	PC	MOS+'σ'	838	33			19" LCD	varying	55%
EPFL/PoliMI	ACR-HR	Raw	40	34	24-40		19"/30" LCD	4-8 H_p	
IVC-1080i	ACR-HR	Raw	29	28			37" LCD	3 H_s	
IVC-RoI	ACR	Raw	25	25			19" CRT	6 H_p	
IVP	ACR	DMOS+σ	42	35	20-38	26%	65" PDP	3 H_p	69%
LIVE(V)	ACR	DMOS+σ	38	29	students		CRT		37%
MMSP-3D(V)	ACR	Raw	20	17	24-37	30%	46" LCD	3 H_s	
MMSP-SVD	PC	Raw	16	16	μ=28	31%	30" LCD	2-3 H_p	
NYU-1	ACR	Raw	22	16-22	students		14" LCD		
NYU-2	ACR-HR	Raw	31	15		16%			
NYU-3	ACR	MOS+σ	33	15	21-33	18%			
NYU-PL	SSCQS	MOS+σ	32	32	students		17" LCD	4-6 H_p	
VQEG-FR	DSCQS	DMOS+σ	287	61-147			19" CRT	5 H_s	79%
VQEG-HD	ACR-HR	Raw	120	24			24-47" LCD	3 H_p	78%

Method — Subjective testing method used (refer to [1], [2] for details). PC: paired comparison.
Data — Type of data available: raw scores, MOS/DMOS, standard deviation (σ) or similar.
Ratings — Average number of valid subjective ratings per image/video.
Female — Percentage of female subjects.
Distance — Viewing distance as a multiple of picture height (H_p) or screen height (H_s).
PSNR — Approximate correlation between PSNR and MOS (where provided).

- *Poly@NYU Video Quality Databases*. Three separate but related tests [41]–[44] using videos with different frame rates and quantization parameters.
- *Poly@NYU Packet Loss (PL) Database* [41], [45]. Small database on the impact of packet loss in H.264 videos. Test clips are only 2 seconds long.
- *VQEG FR-TV Phase I Database* [46], [47]. The oldest public quality database (interestingly it came out several years before the first image quality database). Consequently, test conditions focus on MPEG-2 compression and transmission and even include some analog distortions.
- *VQEG HDTV Database* [48]. Test conditions include MPEG-2 and H.264 compression as well as different types of network impairments. 5 of the 6 sets in the HDTV test are being released via the Consumer Digital Video Library (CDVL) [49]; the sixth set is not public. Only the data from the 5 public sets are used in this paper.

The videos in these databases are around 10 seconds long, with the exception of the NYU-PL database. Four databases in this list include the encoded bitstreams, the rest only provide the decoded video frames, as indicated in Table I.

D. Other Databases

There are a number of other databases of relevance to image and video quality researchers. Some annotated databases became available after the work on this paper was completed and are absent from the analysis presented here. A new database for mobile video quality assessment is described in another paper in this special issue [50]. IRCCyN/IVC also added several recent video databases to its web site.[1]

Furthermore, some eye tracking experiments were designed specifically with quality assessment in mind, by including compressed versions of the stimuli [51] or focusing on the quality scoring task [52].

Finally, uncompressed source content is extremely useful and valuable for many areas of image and video processing. The *Consumer Digital Video Library (CDVL)* [49] is one such collection. For 3D content, the *Mobile 3DTV* project [53] provides a number of stereo and multiview videos; the *RMIT 3DV* library [54] contains over 30 HD stereo sequences.

III. ANALYSIS

For the analyses in this section, we focus on three aspects of annotated databases: source content (i.e., reference images/videos), test material (i.e., the samples processed by various test conditions, a.k.a. hypothetical reference circuits), and subjective ratings. We propose a number of quantitative criteria to characterize and visualize these aspects to facilitate comparisons across databases.

A. Source Content

To characterize the source images and videos in each database along the dimensions of color, space, and time, we compute the following parameters:

[1]http://www.irccyn.ec-nantes.fr/spip.php?article491

- *Spatial information* (SI) as an indicator of edge energy [55]. Let s_h and s_v denote the gray-scale images filtered with horizontal and vertical Sobel kernels, respectively. $s_r = \sqrt{s_v + s_h}$ then represents the edge magnitude at every pixel. The SI value used here is the root mean square of the edge magnitude over the image or frame:

$$\mathrm{SI} = \sqrt{\frac{L}{1080}}\sqrt{\sum \frac{s_r^2}{P}}, \tag{1}$$

where P is the number of pixels in the filtered image. The normalization factor $\sqrt{L/1080}$ (L is the number of lines, i.e., vertical resolution) is a somewhat crude but necessary step to reduce the scale/resolution-dependence of SI. SI is computed on the luminance, which (for images or videos that are in RGB format) is obtained with the following conversion formula:

$$Y = 0.299R + 0.587G + 0.114B. \tag{2}$$

- *Colorfulness* (CF) as a perceptual indicator of the variety and intensity of colors in the image. Using $rg = R - G$ and $yb = 0.5(R + G) - B$ as a simple opponent color space, colorfulness is defined as [56]:

$$\mathrm{CF} = \sqrt{\sigma_{rg}^2 + \sigma_{by}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{by}^2}. \tag{3}$$

- *Motion vectors* (MV) as an indicator of motion energy for video.[2] Let $\mathbf{v}$ be the motion vector of a block between two consecutive frames. MV is the normalized root mean square of the motion vector magnitudes across all blocks and frames:

$$\mathrm{MV} = \frac{f}{L}\sqrt{\sum \frac{|\mathbf{v}|^2}{M}}, \tag{4}$$

where M is the number of motion vectors (blocks) in the video. Simple normalization by the number of lines per frame L and the time interval between frames $1/f$ ensures MV remains comparable across different resolutions and frame rates. The MVTools 2.5.11.3 plugin[3] was used with AVISynth 2.58 and VirtualDub 1.9.11 for motion vector estimation (function "MAnalyse", default settings, 8×8 pixels block size).

For video, SI and CF values are averaged over all frames. Chroma upsampling and color conversion of video frames was done using functions from the Intel Integrated Performance Primitives (IPP) 5.3.[4]

The raw SI, CF, and MV values for each database are shown in Fig. 1. As can be seen from the plots, there are quite dramatic differences in the distribution of source content along these dimensions across databases.

Using the above source content characteristics (let's call them C_i, where $C_1 = \mathrm{SI}$, $C_2 = \mathrm{CF}$, $C_3 = \mathrm{MV}$), we want to assess numerically how well the space of all possible sources is covered by a given database. We propose the following criteria for each relevant dimension i:

- *Range* of source characteristic C_i over all sources in the database:

$$R_i^{\mathrm{SRC}} = \frac{\max(C_i) - \min(C_i)}{C_i^{\max}}, \tag{5}$$

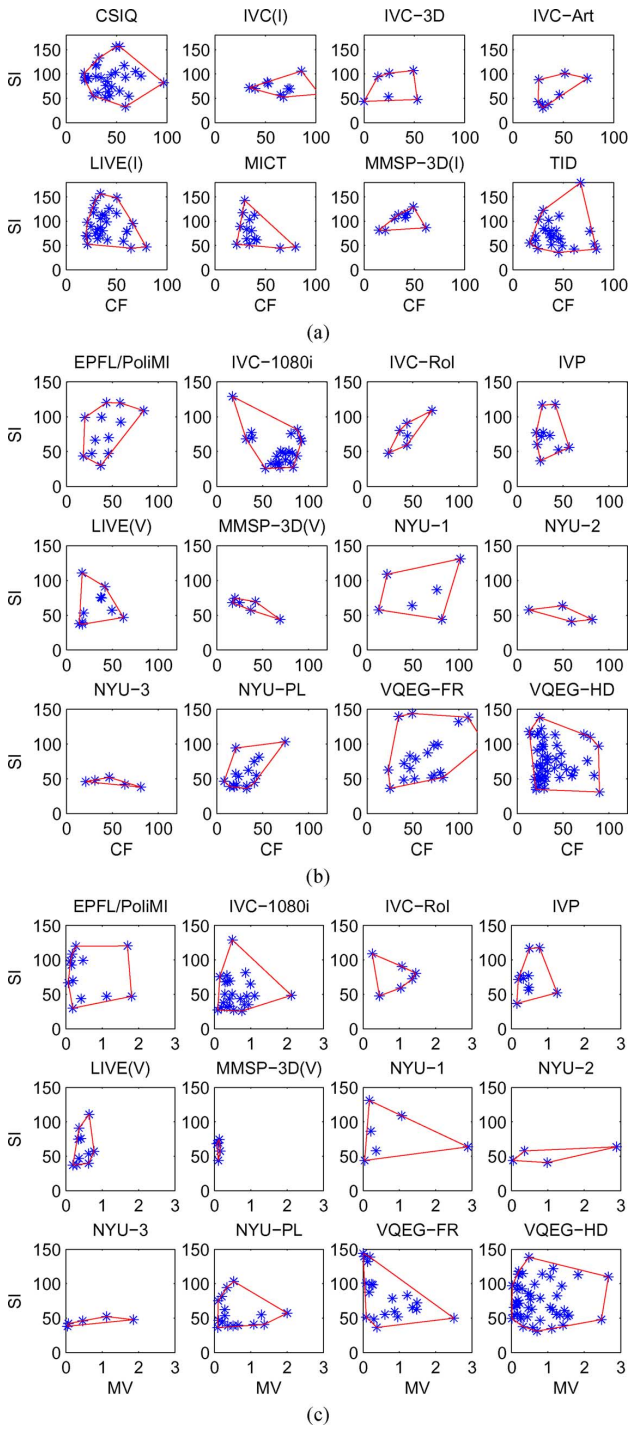

Fig. 1. Spatial information (SI) versus colorfulness (CF) and motion vectors (MV), respectively, and corresponding convex hulls (red lines). (a) $\mathrm{SI} \times \mathrm{CF}$ for color image databases. (b) $\mathrm{SI} \times \mathrm{CF}$ for video databases. (c) $\mathrm{SI} \times \mathrm{MV}$ for video databases.

[2]Temporal information (TI) is another commonly used indicator of motion energy; however, it is highly correlated with spatial complexity and thus less suitable for separate motion classification.

[3]http://avisynth.org.ru/mvtools/mvtools2.html

[4]http://software.intel.com/en-us/articles/intel-ipp/

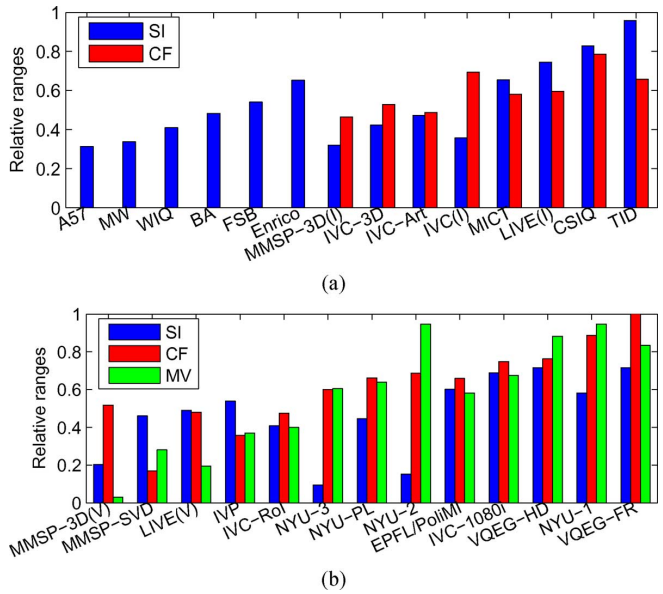

Fig. 2. Relative ranges R_i^{SRC} of source characteristics SI, CF, and MV. (a) Image databases. (b) Video databases.

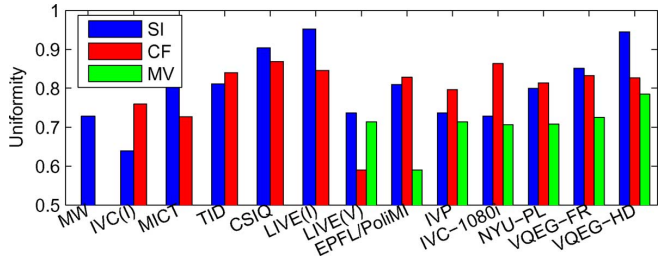

Fig. 3. Uniformity U_i^{SRC} of source characteristics SI, CF, and MV for databases with 10 or more sources.

where $C_i^{\max}$ specifies some maximum value for the given dimension, so that the maximum range is approximately 1 (we use $C_1^{\max} = 150$, $C_1^{\max} = 100$ and $C_3^{\max} = 3$).

- *Uniformity* of coverage. We compute this as the entropy of the B-bin histogram of C_i over all sources in the database.

$$U_i^{\mathrm{SRC}} = -\sum_{k=1}^{B} p_k \log_B p_k, \tag{6}$$

where p_k is the normalized number of sources in bin k. We choose $B = 10$. The entropy is highest ($U = 1$) for completely uniform distributions.

These criteria are plotted in Figs. 2 and 3, quantifying the intra- and inter-database differences in source content characteristics. Because the uniformity criterion is not meaningful when there are few data points, only those databases with at least 10 sources are shown. Overall, most databases fare rather poorly in terms of both criteria.

Since we have up to three dimensions of these basic 1-D criteria, it is helpful to define an additional criterion that expresses the coverage of the 2- or 3-dimensional space of source characteristics. We propose the following:

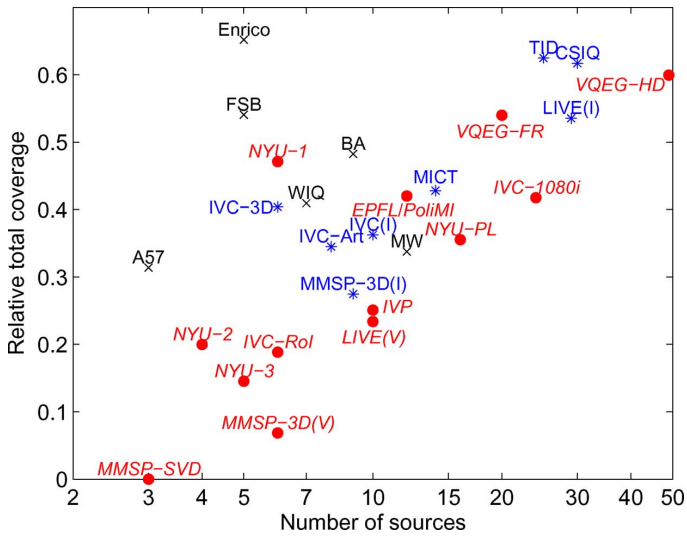

Fig. 4. Relative total coverage T versus number of sources (grayscale databases: black/crosses; color image databases: blue/stars; video databases: red/dots/italic).

- Relative *total coverage* T based on the convex envelope/hull of the set of points in normalized $C_i/C_i^{\max}$ space. For gray-scale images,

$$T = \frac{R_1^{\mathrm{SRC}}}{C_1^{\max}}. \tag{7}$$

For color images, T is the square root of the area of the convex envelope of all points in normalized SI$\times$CF space. For video, T is the cube root of the volume of the convex hull of all points in normalized MV $\times$ SI $\times$ CF space.

The relative total coverage T for all databases is plotted as a function of the number of sources in Fig. 4. As can be expected, a larger number of sources generally improves coverage, even though some databases with fewer sources are surprisingly "efficient" in terms of the range of source characteristics (e.g., Enrico, NYU-1). Overall, the databases with the most content variety manage to cover about 50–60% of the possible range in each dimension, while the bottom end lies around 10–20%.

Finally, we count the scene changes (both cuts and blends) for each source video. Only a handful of databases have sources with scene changes:

- IVC-1080i: 2/1/1 sources (out of 24) with 1/3/4 scene changes, respectively.
- IVC-RoI: 1 source (out of 6) with 1 scene change.
- IVP: 2/1 sources (out of 10) with 1/5 scene changes.
- VQEG-FR: 3/1 sources (out of 20) with 1/3 scene changes.
- VQEG-HD: 5/13/2/2/2/1 sources (out of 49) with 1/2/3/4/5/7 scene changes.

As can be seen, some databases include a few sources with few scene changes. The notable exception is the VQEG-HD database, where over half the source videos contain scene changes.

B. Test Material

To analyze the processed test samples, we compute the peak signal-to-noise ratio (PSNR) for all test images and videos as a rough indicator of the overall range of distortions in each

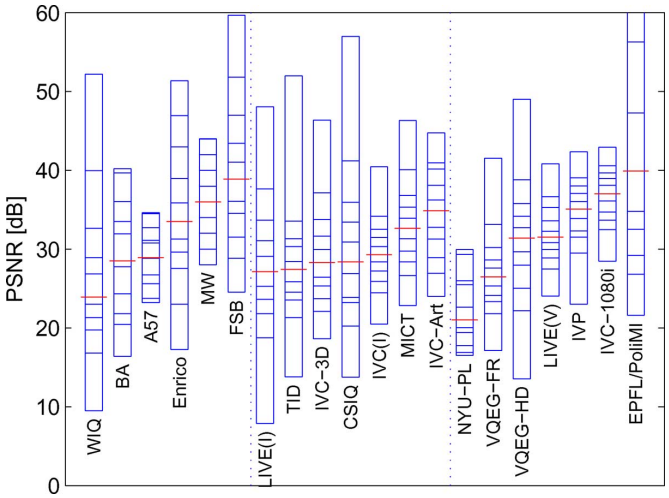

Fig. 5. PSNR density histograms for all databases, sorted by their median (denoted by a wider red line) within each category. Each individual rectangle represents a bin containing 10% of the data points.

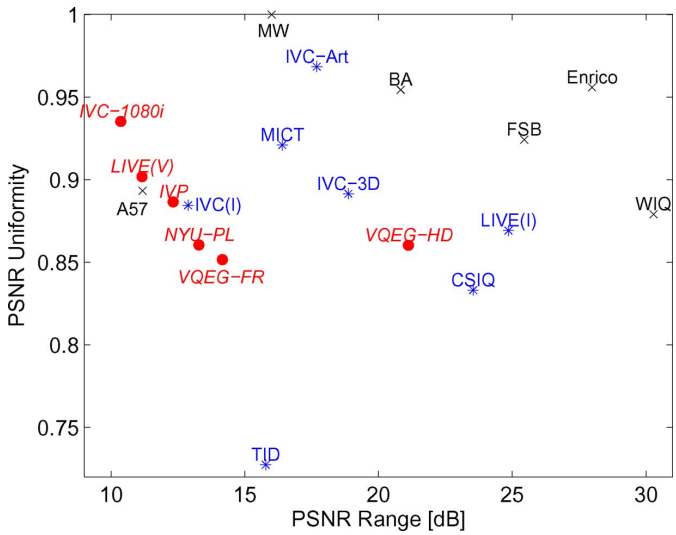

Fig. 6. Uniformity U^{PSNR} versus range R^{PSNR} of PSNR distributions.

database.[5] We use the simplest possible method to compute PSNR here, from the luminance channel only (see (2) above for RGB-to-luminance conversion where necessary). No spatial/temporal alignment or brightness/contrast equalization is performed.

Fig. 5 shows density histograms of PSNR for all databases. The median PSNR values cover a relatively wide range from a low of 20 dB (NYU-PL) to 40 dB (FSB, EPFL/PoliMI[6]). The concentration of distortions around certain PSNR levels is also visible.

Similar to what we defined for characterizing the source content above, we use the following criteria for the test material and its distortions:

- *Range* of PSNR over all test samples in the database:

$$R^{\mathrm{PSNR}} = P^{\mathrm{PSNR}}_{100-n} - P^{\mathrm{PSNR}}_{n}, \tag{8}$$

  where P^{PSNR}_{n} is the n-th percentile of PSNR values in ascending order, with linear interpolation between ranks where necessary. We choose $n = 5$.[7]
- *Uniformity* of coverage U^{PSNR}. We compute this again as the entropy of the 10-bin histogram of PSNR values over all test samples in the database, analogous to (6).

Fig. 6 shows the range and uniformity of PSNR for all databases. WIQ, LIVE (image), and VQEG-HD databases have the largest ranges in their categories. Uniformity is somewhat mixed, with few databases doing very well.

[5]Several video databases are absent from the PSNR analysis in this section: IVC-RoI does not include the source videos; MMSP-3D does not introduce any video distortions; for MMSP-SVD and the NYU databases, PSNR is ill-defined because of the scalable coding framework.

[6]The only distortions in the EPFL/PoliMI database are packet losses (the reference clips are already compressed), which in some cases affect only a few frames, leading to very high average PSNR.

[7]When comparing PSNR ranges across databases, it is important to note that images or video with very high PSNR (from 40–50 dB onwards) may not exhibit visible distortions.

C. Subjective Ratings

The main aspects considered in the analysis of the subjective ratings here are the distributions of the mean (MOS) and standard deviation σ of the subjective ratings, as these are indicative of the quality range of the test material and the precision of the results.[8] To make them comparable for all experiments, all rating scales are normalized by linear transformation to a common 0–100 scale.

Fig. 7 shows the standard deviations as a function of MOS of each database.[9] The differences in the distributions of MOS and standard deviation between databases are evident, especially for video. At the same time, the plots highlight two interesting features common to many databases:

- The standard deviation is typically highest around the middle of the MOS range and decreases towards the ends of the scale. This inverted-U shape can be observed for most databases and subjective experiments, independently of the rating scale used. As we demonstrated previously in [57], this is due largely to the clipping of ratings towards the ends of the scale.
- The data points fall on a kind of arched grid pattern for databases using discrete 5-point scales, because of the limited number of possible MOS-σ combinations for such a coarse scale. This is most visible here for some of the image databases (Enrico, FSB, MW, IVC, MICT) as well as the VQEG-HD database.

We use the following criteria for quantifying the characteristics of subjective ratings in a database:

- *Range* of MOS over all test images/videos in the database, again based on percentiles:

$$R^{\mathrm{MOS}} = P^{\mathrm{MOS}}_{100-n} - P^{\mathrm{MOS}}_{n}, \tag{9}$$

[8]MMSP-SVD is absent from the analysis in this section because it only provides paired-comparison data.

[9]The A57, LIVE (image), and TID databases are excluded from these plots because their standard deviations could not be confirmed.

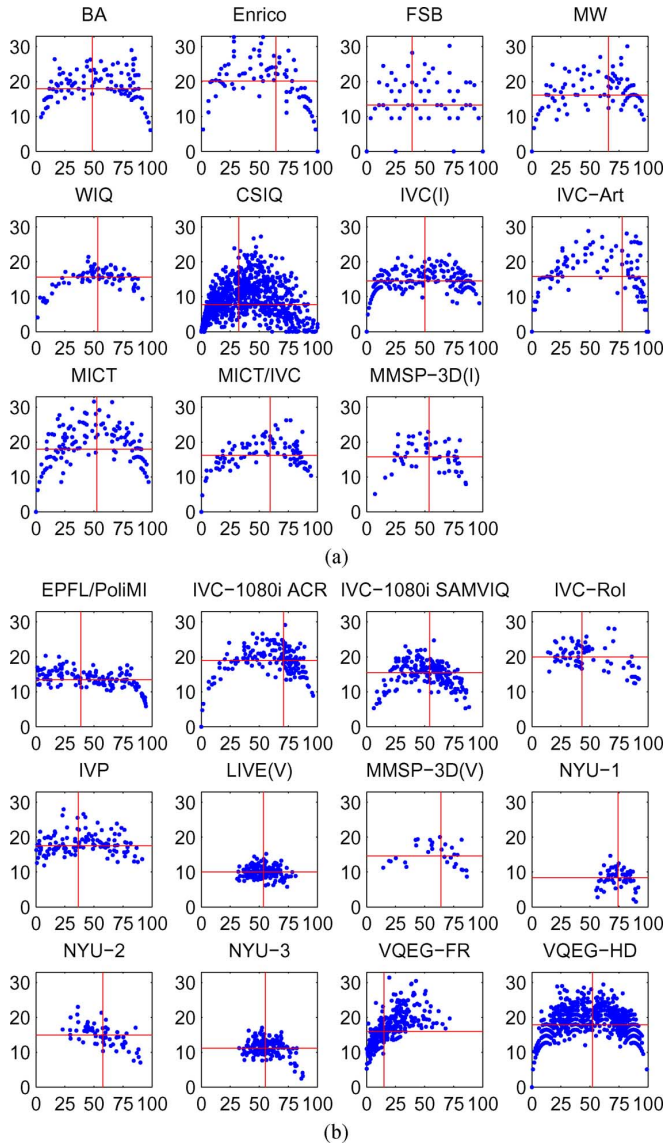

Fig. 7. MOS versus standard deviation of subjective ratings. The red lines mark the respective medians. (a) Image databases. (b) Video databases.

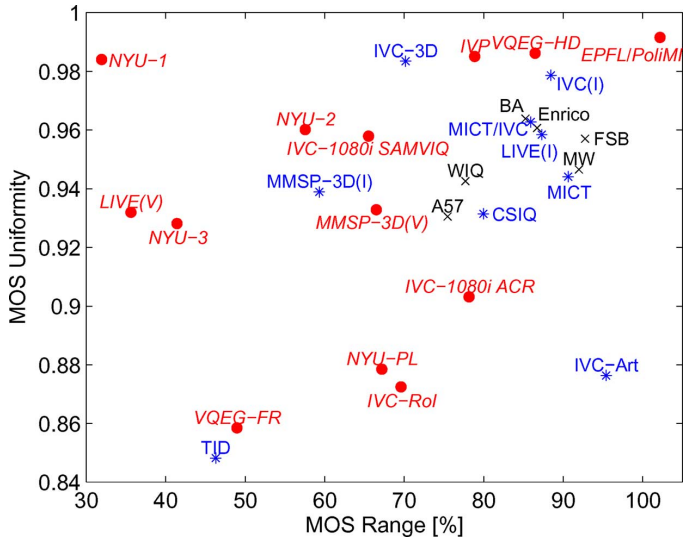

Fig. 8. Uniformity U^{MOS} versus Range R^{MOS} of MOS distributions.

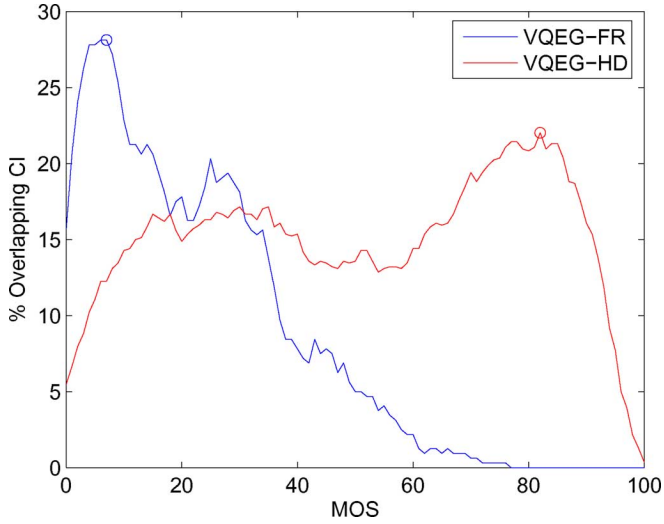

Fig. 9. Sample histograms of overlapping confidence intervals for VQEG-FR and VQEG-HD databases. The circles denote the maxima used for the definition of discriminability D.

where P_n^{MOS} is the n-th percentile of MOS values in ascending order, with linear interpolation between ranks where necessary. We choose $n = 5$.

- *Uniformity* of coverage, U^{MOS}. We compute this again as the entropy of the 10-bin histogram of MOS values in the database, analogous to (6). Uniformity is important because perceived quality levels should be more or less equally distributed across the whole range and not emphasize one part of the scale over another.
- *Variability* V of the ratings. It is computed as the median of standard deviations, restricted to those σ with MOS values in the middle 25% of the range (e.g., from 2.5 to 3.5 on the 1–5 scale). This restriction is necessary because of the significant variation of the standard deviation with MOS as discussed above and evidenced by Fig. 7. A small variability is good because it means the average is more "reliable", and confidence intervals are smaller.

Fig. 8 shows uniformity versus range of MOS distributions. While MOS uniformity is generally quite high, especially compared to source or PSNR uniformity, the plot illustrates the difficulty of getting subjects to use the full range of the rating scales: the MOS for several of the databases barely cover half of the available range (e.g., LIVE (video), TID, VQEG-FR, and a few others).

We further define *Discriminability* D based on the maximum percentage of overlapping 95%-confidence intervals (CI), which is computed as follows. We create a histogram, where all bins (bin size is 1% of the MOS scale) within the range $\mathrm{MOS} \pm \mathrm{CI}$ are incremented by 1 for each test sample. After doing this for all MOS values, the bin counts are normalized by the number of test samples in the database. Examples of this histogram are shown in Fig. 9. D is defined as 1 minus the relative number of entries in the largest bin. It is an indicator of how well subjects were able to distinguish individual test images/videos across the database. D responds to range and uniformity of MOS coverage, as well as MOS variability.

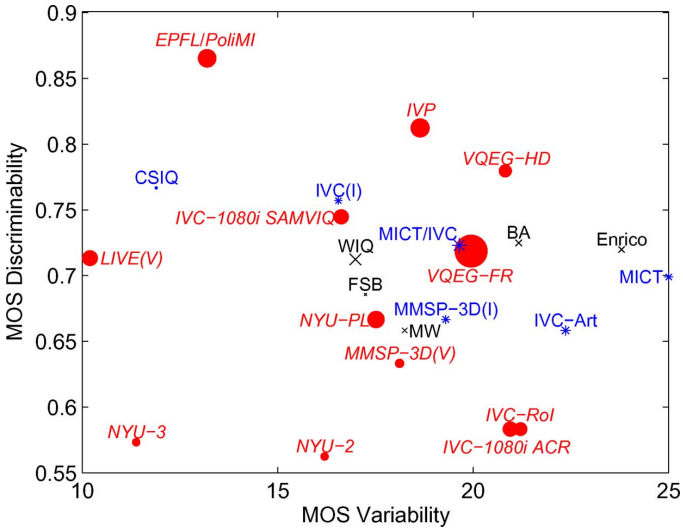

Fig. 10. Discriminability D versus Variability V of subjective ratings. Marker radius is proportional to the number of valid subjective ratings.

Fig. 10 shows the Discriminability D as a function of Variability V.[10] The first thing to note is that there is no obvious correlation between these two parameters, or with the number of subjective ratings. Considering Variability alone, most databases are somewhere in the range of 15–25% of the rating scale, with some going as low as 10%. The EPFL/PoliMI database has the highest discriminability, with overlapping confidence intervals for at most 13% of test samples. For the worst cases, discriminability falls below 0.6, indicating that the MOS values of nearly half the test samples are not significantly different.

IV. Discussion

There are many criteria that can be used to assess and compare databases. We proposed numerical expressions for some of them in this paper. These quantitative criteria can also be helpful in designing a new database and subjective experiment, even though not all of them are easy to target or establish before carrying out the actual tests. Also note that optimizing for one or more of these parameters will not necessarily result in a "better" database. Ultimately, the purpose of an image or video quality database should be to retrieve the advantages and disadvantages of all tested quality metrics [25].

While a database is usually designed with a specific application in mind, the larger the number of sources and test conditions, the more aspects of the problem space can be covered. Variety is especially important for the sources, in order to make sure the database is general enough for extension to content that was not part of the actual test. Characteristics to look out for include spatial complexity (textures, patterns), temporal complexity (object and camera motion), color, faces, text, etc. This is where many databases fall short, as Section III-A highlights, both in terms of uniformity and range of the source characteristics that we analyzed. Furthermore, there is quite a bit of overlap in source material between databases, particularly for video, as there is a limited amount of uncompressed video content available in the public domain.

Of course it is difficult to create source content with very specific amounts of say SI or MV and especially combinations thereof, but the data provided here can be used as a guide to determine where content is lacking. For example, there are relatively few clips with high motion content; especially combinations of high MV with high SI are missing. Identifying such content and making it available would help improve future databases.

Likewise, a large variety of test conditions, distortion levels and types is important not only for the purpose of sampling the entire space, but also for obtaining discriminative ratings from the subjects. This area is perhaps the easiest to address, and the data shown in Section III-B attest to this. Note however that the proposed criteria do not account for distortion variety, which is another important aspect; most databases are highly focused on compression and/or transmission, while ignoring the large number of possible distortions relevant in other applications.

Subjective ratings are the most valuable and perhaps also the trickiest part. One aspect we considered is the distribution of MOS values in terms of range and uniformity. For example, one of the common criticisms of the VQEG FRTV-I database is its bias towards the high quality range. As shown in Section III-C, many other databases have similar problems with MOS range, even though uniformity is generally good.

Adhering to a rigorous methodology is important when conducting subjective experiments, but perhaps even more important is documenting the details of the experimental design and execution. Unfortunately, this is an aspect where many of the current databases fall short, as the sparseness of Table II shows. Perhaps these tables and the discussions here can serve as a rough documentation framework for future database releases.

Viewing conditions also matter: Screen type and resolution, room setup, lighting, viewing distance, and other experiment-specific parameters should be documented. However, there are two schools of thought: some believe that well-controlled experimental conditions and strict compliance with ITU recommendations (as can only be achieved in a lab environment) are essential [58], whereas others argue that naturally variable viewing conditions as users experience in their daily life (different screens, viewing distances, light levels, etc.) are preferable to collect realistic MOS data [25], [59]. The TID database is an example of the latter approach.

Subject pre-screening (i.e., vision tests) and post-screening (i.e., removal of suspicious scores) should be performed, and the specific methods applied (e.g., from [1]) should be mentioned. The subjects demographics (age, gender, etc.) can also have an impact on MOS [60]. In many experiments, male university students are the majority, which may affect the ability to draw conclusions about the responses of other population groups. Again, this information should be included with a database. Finally, it should be considered good practice to release the raw subjective ratings rather than just MOS and σ, so that users can do their own verification or further analysis (for example, this would have been useful to confirm the "standard deviations" included with the LIVE (image) and TID databases).

[10]The A57, LIVE (image), and TID databases are excluded from these plots because their standard deviations could not be confirmed.

Since there is no single best database, and because separate databases cannot simply be combined into one, most metric developers train and/or benchmark their algorithms on multiple databases independently (typically maximizing individual correlations or minimizing prediction errors). However, this is prone to biases from various database-specific peculiarities. While there are methods to cross-calibrate multiple quality metrics using a common dataset [61], better approaches to benchmarking using multiple datasets are also needed.

Comparison with the subjective ratings from annotated image or video quality databases has become the standard approach for testing quality metrics; it tries to answer the question, "how accurate is this metric?" However, this is by no means the only question that can be asked, or perhaps even the best. For example, quality metrics can be evaluated and improved by generating specific types and degrees of distortions [62]. An approach based on the premises of software testing was proposed in [63], where the goal is to expose errors rather than demonstrating that the system satisfies certain specifications. Based on this idea, a methodology for systematic stress testing of quality metrics was developed, which attempts to determine whether a quality metric is inaccurate [64]. Since these approaches do not rely on full-fledged subjective tests, they can serve as valuable complements in metric testing and evaluation.

V. Conclusions

More than two dozen annotated image and video quality databases are now available in the public domain. This encouraging development facilitates benchmarking of algorithms (see also [65] for more on this topic) and helps make models more comparable.

In addition to a detailed overview of these databases, this paper proposed several quantitative criteria and analysis methods for source content, test material, and subjective ratings, that allow analytical comparisons of databases.

The list of databases is bound to grow as new applications (e.g., 3D or multi-view video) emerge. The data presented here may be useful for identifying content or test material that is currently missing from the public domain. One type of database that is particularly lacking is audiovisual content, an area which generally deserves more attention. An up-to-date list of links to the various databases is available on the author's home page [3].

Acknowledgment

The author is grateful to all those researchers who created and released the databases mentioned in this article. He would also like to thank D. Demircioğlu for helping with the data collection, M. Pinson for her valuable comments on an early draft of the manuscript, and the reviewers for their helpful suggestions.

References

[1] "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, 2012, ITU-R Rec. BT.500-13.

[2] "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, Switzerland, 2008, ITU-T Rec. P.910.

[3] S. Winkler, "Image and video quality resources," 2012 [Online]. Available: http://stefan.winkler.net/resources.html

[4] D. M. Chandler and S. S. Hemami, "VSNR online supplement," 2007 [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

[5] F. Autrusseau *et al.*, "IVC Watermarking Databases (Broken Arrow, Enrico, Fourier Subband, Meerwald)," 2010 [Online]. Available: http://www.irccyn.ec-nantes.fr/~autrusse/Databases/

[6] E. Marini, F. Autrusseau, P. Le Callet, and P. Campisi, "Evaluation of standard watermarking techniques," in *Proc. SPIE Security and Watermarking of Multimedia Contents*, San Jose, CA, Jan. 2007, vol. 6505.

[7] M. Carosi, V. Pankajakshan, and F. Autrusseau, "Toward a simplified perceptual quality metric for watermarking applications," in *Proc. SPIE Multimedia on Mobile Devices*, San Jose, CA, Jan. 2010, vol. 7542.

[8] U. Engelke, H.-J. Zepernick, and M. Kusuma, "Wireless imaging quality database," 2010 [Online]. Available: http://www.bth.se/tek/rcg.nsf/pages/wiq-db

[9] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Process.: Image Commun.*, vol. 24, no. 7, pp. 525–547, 2009.

[10] E. C. Larson and D. M. Chandler, "Consumer subjective image quality database," 2009 [Online]. Available: http://vision.okstate.edu/index.php?loc=csiq

[11] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, Mar. 2010.

[12] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," 2005 [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/

[13] A. Ninassi, P. Le Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *Proc. SPIE Human Vis. Electron. Imaging*, San Jose, CA, Jan. 2006, vol. 6057.

[14] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "IRCCyN/IVC 3D images database," 2008 [Online]. Available: http://www.irccyn.ec-nantes.fr/spip.php?article876

[15] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP J. Image Video Process.*, 2008, Article ID 659024.

[16] F. Autrusseau and M. Babel, "Subjective quality assessment of LAR coded art images," 2009 [Online]. Available: http://www.irccyn.ec-nantes.fr/~autrusse/Databases/

[17] C. Strauss *et al.*, "Subjective and objective quality evaluation of LAR coded art images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, New York, Jun.-Jul. 28–3, 2009.

[18] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2," 2006 [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm

[19] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[20] Z. M. Parvez Sazzad, Y. Kawayoke, and Y. Horita, "MICT image quality evaluation database," 2008 [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html

[21] S. Tourancheau, F. Autrusseau, Z. M. Parvez Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Proc. Int. Conf. Image Processing (ICIP)*, San Diego, CA, Oct. 12–15, 2008.

[22] L. Goldmann *et al.*, "3D image quality assessment," 2010 [Online]. Available: http://mmspl.epfl.ch/page38842.html

[23] L. Goldmann, F. De Simone, and T. Ebrahimi, "Impact of acquisition distortion on the quality of stereoscopic images," in *Proc. Int. Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, AZ, Jan. 13–15, 2010.

[24] N. Ponomarenko *et al.*, "TID2008 – A database for evaluation of full-reference visual quality assessment metrics," 2008 [Online]. Available: http://www.ponomarenko.info/tid2008.htm

[25] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 – A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, pp. 30–45, 2009.

[26] F. De Simone *et al.*, "EPFL-PoliMI video quality assessment database," 2009 [Online]. Available: http://vqa.como.polimi.it/

[27] F. De Simone *et al.*, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in *Proc. Int. Workshop Quality of Multimedia Experience (QoMEX)*, San Diego, CA, Jul. 29–31, 2009.

[28] S. Péchard, R. Pépion, and P. Le Callet, "IRCCyN IVC 1080i database," 2008 [Online]. Available: http://www.irccyn.ec-nantes.fr/spip.php?article541
[29] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm," in *Proc. Int. Workshop Image Media Quality and its Applicat. (IMQA)*, Kyoto, Japan, Sep. 2008.
[30] F. Kozamernik, P. Sunna, E. Wyckens, and D. I. Pettersen, "Subjective quality of internet video codecs – Phase II evaluations using SAMVIQ," *EBU Tech. Rev.*, no. 301, Jan. 2005.
[31] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "IRCCyN IVC SD RoI database," 2009 [Online]. Available: http://www.irccyn.ec-nantes.fr/spip.php?article551
[32] S. Péchard, R. Pépion, and P. Le Callet, "Region-of-interest intra prediction for H.264/AVC error resilience," in *Proc. Int. Conf. Image Process. (ICIP)*, Cairo, Nov. 7–10, 2009.
[33] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP subjective quality video database," 2011 [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/
[34] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "LIVE video quality database," 2010 [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html
[35] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[36] L. Goldmann *et al.*, "3D video quality assessment," 2010 [Online]. Available: http://mmspl.epfl.ch/page38842.html
[37] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *Proc. SPIE 3D Image Process. (3DIP) and Applicat.*, San Jose, CA, Jan. 17–21, 2010, vol. 7526.
[38] J.-S. Lee *et al.*, "MMSP scalable video database," 2010 [Online]. Available: http://mmspg.epfl.ch/svd
[39] J.-S. Lee *et al.*, "Subjective evaluation of scalable video coding for content distribution," in *Proc. ACM Multimedia*, Firenze, Italy, Oct. 25–29, 2010, pp. 65–72.
[40] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 882–893, Oct. 2011.
[41] Y. Wang *et al.*, "Poly@NYU video quality databases," 2008 [Online]. Available: http://vision.poly.edu/index.html/index.php?n=HomePage.QualityAssessmentDatabase
[42] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang, "Modeling the impact of frame rate on perceptual quality of video," in *Proc. Int. Conf. Image Process. (ICIP)*, San Diego, CA, Oct. 12–15, 2008.
[43] Y.-F. Ou, Z. Ma, and Y. Wang, "A novel quality metric for compressed video considering both frame rate and quantization artifacts," in *Proc. Int. Workshop Video Process. Quality Metrics (VPQM)*, Scottsdale, AZ, Jan. 15–16, 2009.
[44] Y.-F. Ou, Y. Zhou, and Y. Wang, "Perceptual quality of video with frame rate variation: A subjective study," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, Mar. 14–19, 2010.
[45] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Proc. Int. Conf. Image Process. (ICIP)*, San Diego, CA, Oct. 12–15, 2008.
[46] "VQEG FR-TV Phase I database," Video Quality Experts Group (VQEG), 2000 [Online]. Available: ftp://ftp.crc.ca/crc/vqeg/TestSequences/
[47] "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," VQEG, Apr. 2000 [Online]. Available: http://www.vqeg.org/
[48] "Report on the validation of video quality models for high definition video content," VQEG, Jun. 2010 [Online]. Available: http://www.vqeg.org/
[49] 'The Consumer Digital Video Library," CDVL, 2010 [Online]. Available: http://www.cdvl.org/
[50] A. K. Moorthy, L. K. Choi, A. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. XXX–XXX, Oct. 2012.
[51] U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Visual attention modeling for subjective image quality databases," in *Proc. Workshop Multimedia Signal Process. (MMSP)*, Rio de Janeiro, Brazil, Oct. 5–7, 2009.
[52] J. Redi, H. Liu, R. Zunino, and I. Heynderickx, "Interactions of visual attention and quality perception," in *Proc. SPIE Human Vis. Electron. Imaging*, San Jose, CA, Jan. 24–27, 2011.
[53] A. Smolic, G. Tech, and H. Brust, "Mobile 3DTV Stereo-Video Content," [Online]. Available: http://sp.cs.tut.fi/mobile3dtv/stereo-video/
[54] E. Cheng, P. Burton, J. Burton, A. Joseski, and I. Burnett, "RMIT3DV: Pre-announcement of a creative commons uncompressed HD 3D video database," in *Proc. Int. Workshop Quality of Multimedia Experience (QoMEX)*, Yarra Valley, Australia, Jul. 2012.
[55] "Digital transport of one-way video signals – Parameters for objective performance assessment," American National Standards Institute, New York, 1996, ANSI T1.801.03.
[56] D. Hasler and S. Süsstrunk, "Measuring colourfulness in natural images," in *Proc. SPIE Human Vis. Electron. Imag.*, Santa Clara, CA, Jan. 21–24, 2003, vol. 5007, pp. 87–95.
[57] S. Winkler, "On the properties of subjective ratings in video quality experiments," in *Proc. Int. Workshop Quality of Multimedia Experience (QoMEX)*, San Diego, CA, Jul. 29–31, 2009.
[58] C. Keimel, T. Oelbaum, and K. Diepold, "Improving the verification process of video quality metrics," in *Proc. Int. Workshop Quality of Multimedia Experience (QoMEX)*, San Diego, CA, Jul. 2009, pp. 121–126.
[59] N. Staelens *et al.*, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *IEEE Trans. Broadcasting*, vol. 56, no. 4, pp. 458–466, Dec. 2010.
[60] S. Jumisko-Pyykkö and J. Häkkinen, "Profiles of the evaluators – Impact of psychographic variables on the consumer-oriented quality assessment of mobile television," in *Proc. SPIE Multimedia Mobile Devices*, San Jose, CA, Jan. 28–29, 2008, vol. 6821.
[61] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: New methods from ATIS/T1A1," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 101–107, Feb. 2004.
[62] A. C. Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1261–1273, Aug. 2008.
[63] F. M. Ciaramello and A. Reibman, "Supplemental subjective testing to evaluate the performance of image and video quality estimators," in *Proc. SPIE Human Vis. Electron. Imag.*, San Jose, CA, Jan. 2011, vol. 7865.
[64] F. M. Ciaramello and A. Reibman, "Systematic stress testing of image quality estimators," in *Proc. Int. Conf. Image Process. (ICIP)*, Brussels, Belgium, Sep. 11–14, 2011.
[65] R. C. Streijl, S. Winkler, and D. S. Hands, "Perceptual quality measurement – Towards a more efficient process for validating objective models," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 136–140, Jul. 2010.

Stefan Winkler is Principal Scientist and Director of the Interactive Digital Media Program at the University of Illinois' Advanced Digital Sciences Center (ADSC) in Singapore. He also serves as Scientific Advisor to Cheetah Technologies. Prior to that, he co-founded a start-up, worked in several large corporations, and held faculty positions at two universities.

Dr. Winkler has a Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, and an M.Eng./B.Eng. degree from the University of Technology Vienna, Austria. He has published over 80 papers and the book "Digital Video Quality." He is an Associate Editor of the IEEE Transactions on Image Processing and the IEEE Signal Processing Magazine (Standards Column). He has also contributed to video quality standards in VQEG, ITU, ATIS, VSF, and SCTE. His research interests include video processing, computer vision, perception, and human-computer interaction.