Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

MCL-V: A streaming video quality assessment database ☆

Joe Yuchieh Lin [a], Rui Song [b,*], Chi-Hao Wu [a], TsungJung Liu [a], Haiqiang Wang [a], C.-C. Jay Kuo [a]

[a] University of Southern California, Ming Hsieh Department of Electrical Engineering, 3740 McClintock Avenue, Los Angeles, CA, United States
[b] Xidian University, 710071 P.O. Box 103, 2nd Taibai South Road, Xi'an, Shaanxi, China

ARTICLE INFO

Article history:
Received 9 June 2014
Accepted 18 February 2015
Available online 5 March 2015

Keywords:
Video quality
Quality assessment
Quality database
Video streaming
Compression distortion
Video re-sizing
Video coding
Perceptual quality

ABSTRACT

A high-definition video quality assessment (VQA) database that captures two typical video distortion types in video services (namely, "compression" and "compression followed by scaling") is presented in this work. The VQA database, called MCL-V, contains 12 source video clips and 96 distorted video clips with subjective assessment scores. The source video clips are selected from a large pool of public-domain high-definition (HD) video sequences with representative and diversified contents. Both distortion types are perceptually adjusted to yield distinguishable distortion levels. An improved pairwise comparison method is adopted for subjective evaluation to save evaluation time. Several existing image and video quality assessment (IQA and VQA) algorithms are evaluated against the MCL-V database. The MCL-V database is publicly accessible in the link – http://mcl.usc.edu/mcl-v-database/ to facilitate future video quality assessment research of the community.

1. Introduction

The high-definition video broadcasting and streaming services are blooming nowadays. Consumers can enjoy on-demand video services from Netflix, Hulu or Amazon, and watching high-definition (HD) programs becomes the mainstream for video content consumption. According to the report in [1], more than half of US population watches on-line movies or dramas. Specifically, the viewers have increased from 37% in 2010 to 51% in 2013. The watched video programs vary in bit rates and resolutions due to the available bandwidth of their networks. Different sizes of video are transmitted at lower bit rates and up-scaled for display on HDTV (e.g., playing a 720p movie on the 1080p screen). This is common in people's daily life [2], but users' video quality of experience on HD video has not yet been extensively studied in the past.

There are quite a few video quality assessment databases available to the public [3–28]. They were however limited in the following areas [29,30]. First, the source video set is not representative or diversified enough. For example, they do not contain dark scenes, sports scenes, traditional cartoon, and computer animation. The lack of these contents will not provide an extensive evaluation of viewers' experience. Second, the video resolution is low. The resolution of sequences in all VQA databases except [3,8,20,26,28] are lower than 1920 × 1080. Third, the distortion is not complete for the target application. For example, all above-mentioned VQA databases except [20,10,11] do not cover video up-scaling, which is encountered frequently in our daily life. Although the work in [20] includes practical distortion types, it has only three video sources. Being motivated by these observations, we build a new VQA database called MCL-V to address the shortcomings of existing VQA databases. The MCL-V database provides 12 source video clips, 96 distorted video clips and their associated mean opinion scores (MOS). In this paper, we will elaborate on the methodology of building MCL-V including collecting suitable video sources, generating distortions and conducting subjective evaluation.

One key issue in our design is to choose an appropriate subjective test procedure to collect opinion scores. Several subjective test methodologies have been recommended in VQEG [25,31] and ITU [32,33] as shown in Table 1. Since the precision of the final MOS is not improved by adopting the continuous scale [34,35], the discrete scale is adopted in this work for user friendliness. Furthermore, we use an improved pairwise comparison method to make the final MOS more stable and meaningful.

The rest of this paper is organized as follows. Section 2 describes ways to choose representative and diversified reference sequences, to generate practical distortion types and to determine the reasonable distortion levels. Section 3 presents an improved pairwise comparison method for subjective evaluation and elaborates on the process of collecting and normalizing opinion scores in the subjective test. We study the MOS values and analyze the performances of several existing IQA and VQA metrics against the MCL-V database in Section 4. Finally, concluding remarks are

☆ This paper has been recommended for acceptance by Prof. M.T. Sun.
* Corresponding author.
E-mail address: rsong@xidian.edu.cn (R. Song).

http://dx.doi.org/10.1016/j.jvcir.2015.02.012
1047-3203/© 2015 Elsevier Inc. All rights reserved.

given in Section 5. The whole database is publicly available on the USC Media Communication Lab website http://mcl.usc.edu/mcl-v-database/.

2. Construction of MCL-V database

2.1. Source video selection

We selected 12 uncompressed HD video clips as the source sequences. Some sequences are originally in YUV444p or YUV422p, and we converted them into YUV420p using FFMpeg [36] to make all videos included in the MCL-V database be YUV420p at a fixed resolution of progressive 1920 × 1080. The frame rates of the sequences range from 24 fps to 30 fps, and the length of each video is 6 s. Fig. 1 shows all reference videos with a single frame.

The selected sequences are freely available from several sources, including HEVC test sequences [37], TUM dataset [38], CDVL [39], and others [40–42]. They were professionally acquired and recorded in digital form. We select some of them to construct the MCL-V database based on the following two criteria.

First, some prior databases [13,23] contain scenes that are not representative in video applications. For example, there are video clips with a close view on the water surface or the blue sky in the LIVE database [23]. These sequences were used for video coding performance test since they contain specific contents which are difficult to encode. However, they are not common scenes in movies or dramas. We prefer more representative scenes since they can better reveal human visual experience.

Second, the database should have sufficient diversity in terms of several characteristics. We list various characteristics for diversity consideration in Table 2. They are categorized into three groups: (1) high-level video genres, (2) mid-level video semantics and (3) low-level video features. We aim to make the database cover a wide range of characteristics given in the table.

For video genres, we take several new genre types such as animation and sports into account. These video genres have different characteristics from others. For instance, cartoons scenes contain clear edges and simple color components while sports scenes contain fast moving objects with simple background. These videos are commonly seen in applications and should be included in the MCL-V database.

For video semantics, we consider factors that will have a great impact on human visual perception. For example, while other databases usually do not include video scenes with a close-up face, we take this feature into consideration since it is typical in many dramas. In addition, the human face is typically a region of visual salience which attracts human attention.

Table 1
Classification of subjective testing methods.

	Discrete scale	Continuous scale
Single Stimulus	Absolute Category Rating (ACR) [32]	Single Stimulus Continuous Quality Evaluation (SSCQE) [33]
Double Stimulus	Degradation Category Rating (DCR) [32] Comparison Category Rating (CCR) [33]	Double Stimulus Continuous Quality Scale (DSCQS) [33]

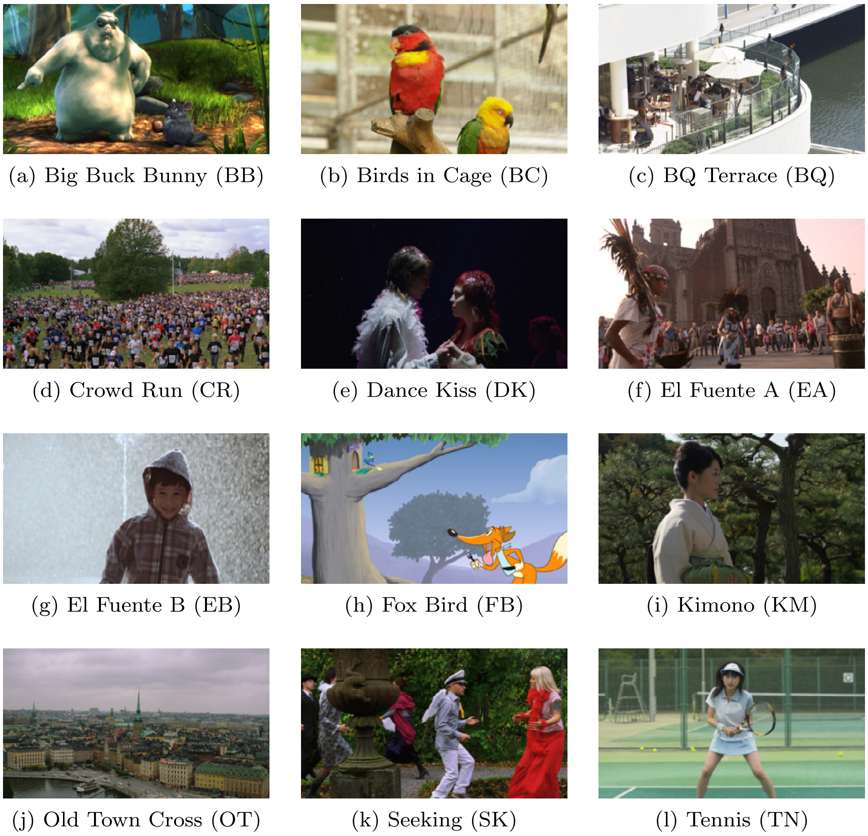
(a) Big Buck Bunny (BB) (b) Birds in Cage (BC) (c) BQ Terrace (BQ)
(d) Crowd Run (CR) (e) Dance Kiss (DK) (f) El Fuente A (EA)
(g) El Fuente B (EB) (h) Fox Bird (FB) (i) Kimono (KM)
(j) Old Town Cross (OT) (k) Seeking (SK) (l) Tennis (TN)

Fig. 1. Selected source video sequences.

Table 2
Video characteristics used in diversity check.

Video genres	Video semantics	Video features
• Cartoon	• Face	• Brightness
• Sports	• People	• Contrast
• Indoor	• Water	• Texture
	• Number of objects	• Motion
	• Salience	• Color Variance
		• Color Richness
		• Sharpness
		• Film Grain
		• Camera motion
		• Scene change

For video features, we examine brightness, contrast, motion, texture and color since these features are related to the level of the video compression distortion. These features also have influence on the visual masking effect. For example, there is no very dark scene or fast-motion scene in existing video quality databases [6,13,23]. As a result, they do not contain representative video clips for horror movies or action films. The diversity of video features can be captured by the Spatial Information (SI) versus the Temporal Information (TI) plot as defined in the ITU-T Recommendation [32]. Eqs. (1) and (2) of SI and TI are shown as follows:

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}, \quad (1)$$

$$TI = max_{time}\{std_{space}[M_n(i,j)]\}, \quad (2)$$

SI is calculated based on the Sobel filter. The nth video frame, F_n, is first filtered with the Sobel filter and taken the standard deviation over space domain. Then, the maximum value along the time is chosen to present SI. TI is based on motion difference. $M_n(i,j)$ is the difference in pixel at the ith row and jth column between F_n and F_{n-1}. TI is computed as the time maximum of the space standard deviation of $M_n(i,j)$. These two indices correspond to the texture and the motion features in Table 2, respectively. As shown in Fig. 2, the 12 video sequences in the MCL-V database are well scattered in the feature space spanned by SI and TI, which demonstrates the diversity of the MCL-V database.

Not all characteristics can be quantitatively measured. We conducted subjective evaluation on the characteristics of video clips to illustrate the diversity of the MCL-V database and show the results in Table 3. The main characteristics are listed from high-to-low levels in the first column while the 12 source sequences are listed in the top row in this table. Each column in the table represents the characteristics of the corresponding source sequence. The subjective evaluation was conducted by a group of professionals. Since there are only a few levels defined for each property, the results can be easily verified and they are quite consistent among viewers. This table shows that the selected source video clips in the MCL-V database well span all characteristics with excellent diversity.

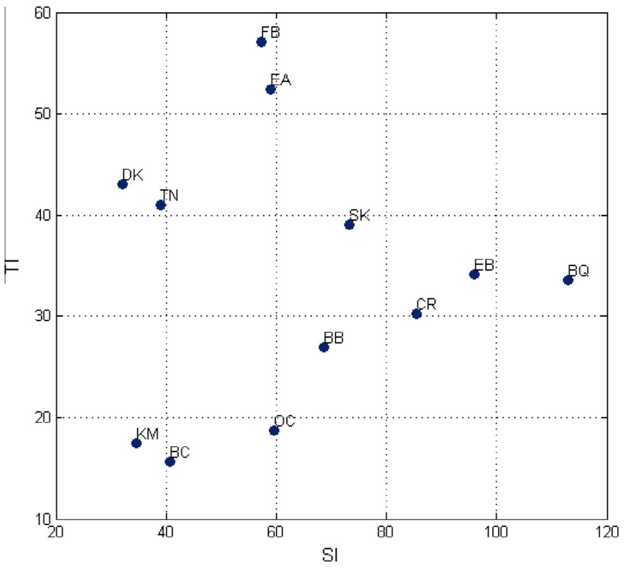

Fig. 2. Plot of the Spatial Information (SI) and the Temporal Information (TI) indices for selected video sequences.

The contents of the 12 source video clips are described below.

- Big Buck Bunny (BB) in Fig. 1a: An animated sequence, where there are two animals in the video, with clear textures and rich backgrounds.
- Birds in Cage (BC) in Fig. 1b: Two colorful birds standing in front of a clean background in a still scene.
- BQ Terrace (BQ) in Fig. 1c: Plenty of vehicles moving on a bridge, and below the bridge are the water. The camera pans in a diagonal direction.
- Crowd Run (CR) in Fig. 1d: A crowd of people running together, with big trees and the blue sky as the background.
- Dance Kiss (DK) in Fig. 1e: People dancing in a very dark room. There are scene changes, and the motions are fast. People will focus on two main characters that kiss in the middle of the scene.
- El Fuente A (EA) in Fig. 1f: Several people in the tribe dancing around a man who is drumming. In addition to fast motions, the scene also contains large portions of ground and sky that are with low gradient.
- El Fuente B (EB) in Fig. 1g: A boy walking in front of a fountain. In another scene, we have a close view to the frontal face of the boy. The water drops in the background make it very difficult for video coding.
- Fox Bird (FB) in Fig. 1h: A cartoon sequence with a fox running rapidly. There are scene changes, and several camera motions are involved.
- Kimono (KM) in Fig. 1i: A woman walking slowly toward the camera in front of the woods. The woman is close to the camera and the face of the women can be seen clearly.
- Old Town Cross (OT) in Fig. 1j: A bird's eye view of an old town with slow camera movements. Except the sky and the buildings, there are no other objects in the scene. Film grain noise can be observed in this video sequence.
- Seeking (SK) in Fig.1k: Several people in different colors moving around.
- Tennis (TN) in Fig. 1l: Girls playing tennis, and running very fast to chase the ball. There is also a scene change in this sequence.

2.2. Distorted video generation

We consider two typical distortion types in video applications.

- H.264/AVC compression.
  H.264/AVC is the most popular video format used in IP-based video streaming. The compression artifact due to lower coding bit rates is one main distortion source.
- Compression followed by scaling (or simply called scaling below).
  The image size has to be scaled when a video clip of a lower resolution is displayed in a display panel of higher resolution. This effect can be simulated via a cascade of operations: *down-sampling, encoding, and then resizing to the original resolution.*

We adopt four distortion levels for each distortion type. Since there are 12 source reference sequences, we have $12 \times 2 \times 4 = 96$ distorted sequences in total.

Table 3
MCL-V source video diversity.

	BB	BC	BQ	CR	DK	EA	EB	FB	KM	OT	SK	TN
Cartoon								√				
CG Animation	√											
Sports												√
Indoor					√							
Scene change					√	√	√	√				√
Camera motion[a]	*P*	*S*	*P*	*M*	*Z*	*P*	*P*	*SPZ*	*P*	*P*	*P*	*P*
Face close-up					√		√		√			
People					√	√	√		√		√	√
Water Surface			√									
Salience	√	√			√		√	√	√			√
Film grain noise										√		
Flat, low gradient area		√				√						
Object number[b]	1	1	2	3	1	2	1	2	1	0	2	1
Brightness	2	3	2	2	1	2	3	3	2	2	3	2
Contrast	3	3	2	3	1	2	3	2	2	1	2	2
Texture (spatial variance)	2	1	2	3	2	2	3	2	3	2	2	1
Motion (temporal variance)	2	1	1	3	3	2	2	3	2	1	2	3
Color variance	1	3	1	3	1	1	1	3	2	1	2	1
Color richness	2	3	1	2	1	1	1	3	2	1	3	2
Sharpness	2	3	2	1	2	2	1	3	3	2	2	1

For high-level video genres, √ indicates the video contains this features, and vice versa. For low-level video features, the number represents the level of the feature, where 1, 2 and 3 mean low, medium and high, respectively.

[a] Camera motion types: *S* for still, *P* for pan, *Z* for zoom, *M* for irregular movements.

[b] Object number: 0, 1, 2 and 3 indicate no main object, one, a few and many objects, respectively.

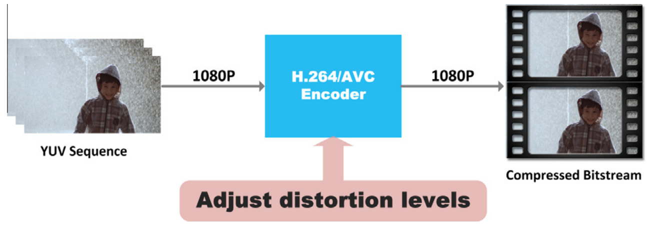

Fig. 3. The process of generating distorted video contents with an H.264/AVC codec.

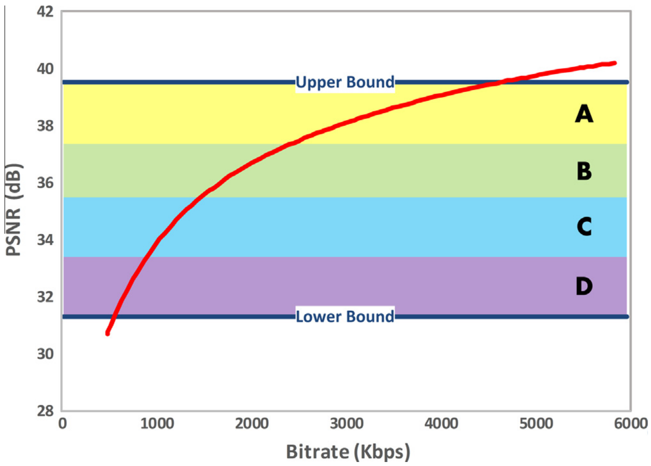

Fig. 4. The process of selecting compression-distorted video clips with four distortion levels.

We used x264 [43] as the encoder to generate compressed video files. Rate control was enabled with a variable bit rate, and a two-pass encoding scheme was used to ensure consistent perceptual quality frame by frame so that viewers can determine the opinion reasonably. At most two B frames are allowed between an I and a P frames. Both the input and the output video resolutions are kept at 1080p as shown in Fig. 3. The distortion levels are controlled by the target bit rates. Since we select a wide variety of video sequences, the bit rate range is from 0.2 Mbps to 10 Mbps.

Since the bit rates depend on video contents, we used the following method to subjectively select distinguishable levels. First, we generated 300 compressed sequences with different bit rates in the above range and drew a plot of “the PSNR value versus the bit rate” as shown in Fig. 4. Although the PSNR value could be used as an auxiliary tool, We do not rely on PSNR to determine perceptual quality. In this bit rate range, there is a region where coded video quality is no distinguishable any longer as the bit rate increases. We also set up a lower bound in the sense that the quality of video clip will not be acceptable if the bit rate is lower than this bound. The perceptual upper and lower bounds are plotted as two solid horizontal lines in Fig. 4. We generated 300 clips for selection within the interval, and divided them into four regions with respect to the PSNR value – A, B, C and D. Finally, we choose four suitable distortion levels (namely, one from each region) based on subjective visual experience.

To generate scaling-distorted video files, we follow the process as depicted in Fig. 5. First, all video sequences are converted to 720p before compression. The down-sampling process is achieved by using the Lanczos algorithm so as to preserve as many details as possible. Different video players may have different settings in video resizing. To make a controllable environment, we choose the bilinear interpolation as the up-sampling algorithm. The format conversion is done by FFmpeg [36]. In the subjective test, we play up-sampled YUV sequences. The distortion levels are adjusted in the compression step, which is the same methodology as before.

3. Subjective video quality assessment

3.1. Subjective assessment methodology

Quite a few subjective test methods for multimedia applications have been recommended by VQEG [25,31] and ITU [32,33]. There are various discrete scoring methods, for example, five score levels in DCR [32] and seven score levels in CCR [33]. When the number of choices increases, it becomes more difficult to get consistent and stable scores across multiple assessors. That is, the same choice made by a different person may have a different meaning.

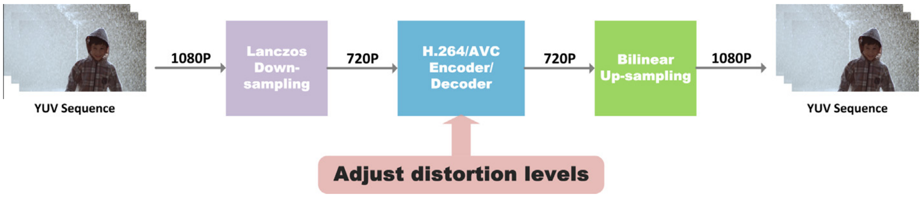

Fig. 5. The process of generating scaling-distorted video clips.

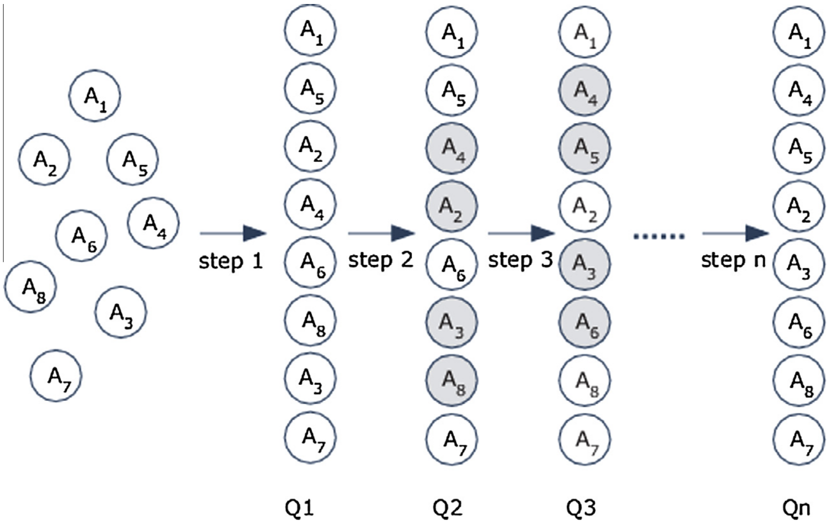

Fig. 6. Illustration of a simplified pairwise comparison process.

Sometimes, the decision of the same person may also vary along the test time. To mitigate these problems, we adopt the pairwise comparison method in the subjective test.

Video clips of the same source but with a different distortion level were selected to form a pair for comparison. An assessor was only asked to decide which video has better quality out of the pair. The objective of a sequence of pairwise comparisons by the same assessor is to create an ordered list of multiple distorted video sequences according to the perceptual quality. The shortcoming of a straightforward pairwise comparison method is its long assessment time. For example, if one attempts to compare the quality of N samples, the total number of an exhaustive pairwise comparison is C_2^N. Several methods were proposed to lower the complexity of the pairwise comparison method, *e.g.*, [44–50]. Here, we propose another simplification method as illustrated in Fig. 6, where each circle represents one distorted sequence. The basic idea is sketched below.

It is desirable to get a good initial list for pairwise comparison. The distorted sequences were first sorted by visual inspection. When the two sequences are far from each other in the queue, it means the visual quality gap between them is obvious. This initialization process is illustrated in Step 1, which is used to generate a rough sorted list of all distorted video sequences for the initialization purpose at a low complexity. Specifically, we ask a small number of professionals to participate in the subjective evaluation with the ACR [32] to achieve this goal. The sorted list result is shown in Q_1, where A_1 and A_7 denote sequences of the best and worst quality, respectively.

After the initialization, all assessors are invited to participate in the subjective test. When the distance of two distorted sequences in the ordered list is longer, their quality difference is more obvious. Thus, each assessor is asked to conduct pairwise comparison of adjacent nodes only. In the given example, if the assessor prefer A_4 to A_2, then A_4 and A_2 are swapped. Furthermore, A_8 and A_3 are swapped similarly. After this round, the assessor is led to a new ordered list denoted by Q_2. With Q_2, the four new adjacent pairs $(A_4, A_5), (A_2, A_6), (A_6, A_3)$, and (A_8, A_7) will be compared by the assessor, and the assessors decision will create Q3. The process is repeated for the same assessor until no further swap is needed. A comparison record matrix is used to record whether any pair of nodes has been compared or not. If two adjacent nodes have been compared by this assessor once, no further comparison will be conducted. All adjacent nodes in the final ordered list, Q_n, will be compared by the same assessor once, and the sequence in the list reflects the preference of this assessor. A preference matrix for the nth assessor, denoted by P_n, can be created accordingly.

By aggregating the preference matrices of multiple assessors, we get the group preference matrix, M. Here, we use the Bradley-Terry model [51–53] to derive the final absolute scale score from the group preference matrix. Note that the Bradley-Terry (BT) model and the Thurstone–Mosteller (T–M) model are two well-known models to convert pair comparison data to psychophysical scale values for all stimuli. To verify its accuracy, we compute the point score, as defined in [54] and compare them in Fig. 7, where the horizontal axis is the point score and the vertical axis is the absolute scale number obtained by using the Bradley-Terry model. We see that the two results are very consistent. The Pearson Correlation Coefficient (PCC) between them is 0.9961. The absolute scale score can also be derived by using the Morrisey Gulliksen incomplete matrix solution [55,56].

3.2. Test setting and procedure

The assessors are seated in a controlled environment to assess the quality of video. The view distance is strictly kept in 2 m (3.2 times of the picture height), from the center of the monitor to

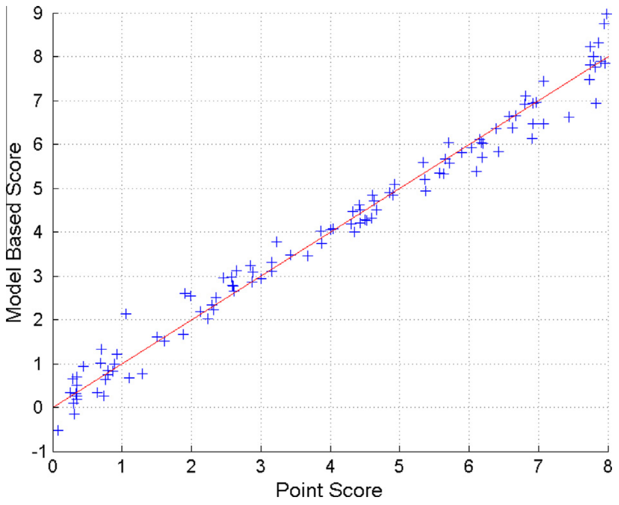

Fig. 7. Correlation between the point score and the absolute scale number calculated based on the Bradley-Terry Model.

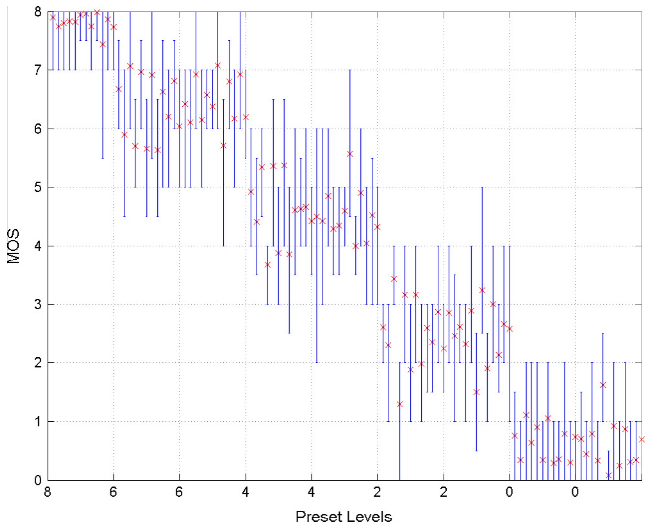

Fig. 8. Sorted Mean Opinion Scores with the 95% confidence interval, where the red cross is the mean and the blue line indicates the stand deviation range between −1 and 1.

Table 4
Comparison of the mean and the variance of opinion scores for compression-distorted Crowd Run and Kimono sequences in MCL-V.

	Mean		Variance	
Level	Crowd run	Kimono	Crowd run	Kimono
Good	6.91	6.92	0.25	0.22
Fair	5.38	4.85	0.51	0.38
Poor	3.16	2.61	0.41	0.16
Bad	1.05	0.80	0.39	0.35

Table 5
Comparison of the MOS values of the compression- and scaling-distorted Dance Kiss and Fox Bird sequences.

	Dance Kiss		Fox Bird	
Level	Compression	Scaling	Compression	Scaling
Good	6.63	6.20	6.58	6.38
Fair	4.61	4.62	4.34	4.59
Poor	2.59	2.35	2.88	1.50
Bad	0.35	0.79	1.61	0.07

the seat. The videos are displayed on the HDTV, LG 47LW5600, with native 1920 × 1080 resolution, thorough this work.

The total number of assessors is 45 consisting of 13 females and 32 males. Their age is distributed from 20 to 40. Some of the assessors are PhD students in image processing field. Others are naive and inexperienced with the topic of video quality assessment. The assessors are confirmed verbally with sound or corrected vision.

Before each test session, a lesson is offered to assessors on how to provide their opinion scores. The training session consists of two parts. For the first part, a 5-min video is played with various video quality. In the second part, assessors learn to see the difference in video quality and the way to operate the software. After the training lesson, assessors will see the notification on the screen and start their test session.

The subjective test is conducted based on the modified pairwise comparison. The software is written in Python. Video clips of the same source but with a different distortion level were selected to form a pair for comparison. They were randomly ordered and played one by one with a 3-s break in-between. An assessor was given three choices: “the first one is better”, “the second one is better”, or “no difference”. Eight video sets were tested at each session and 45 sessions were conducted. One video set includes all distorted video clips from the same video content. We collected 32 opinion scores for each video set. Most assessors have no prior experience in video coding. The test time and each decision made by every assessor were recorded for outlier detection and score conversion. The duration of a test session ranges from 20 to 30 min.

4. Analysis of subjective opinion scores

The collected opinion scores are processed according to the ITU recommendation [33]. The screening of possible outlier subjects is done by following that in [54]. That is, the highest 10% and lowest 10% of the point scores are discarded. The final MOS values with the 95% confidence interval sorted along the decreasing preset level are shown in Fig. 8. We see that the MOS values range from 0 to 8. The mean and the standard deviation of assessors’ scores for each distorted video file are provided in the MCL-V database. In this section, we discuss how video properties affect these scores in the following two scenarios.

First, we compare the mean and the variance of opinion scores for two compressed sequences in Table 4: Crowd Run and Kimono. As shown in the table, the variance of Kimono is significantly lower than that of Crowd Run. This can be explained as follows. Human has a clear visual attention region in Kimono, which is the Japanese lady in the scene. In contrast, there is no clear visual attention region in Crowd Run. As a result, viewers’ opinions are more diverse for Crowd Run.

Next, we compare the MOS values for two compression-distorted and scaling-distorted video clips, Dance Kiss and Fox Bird, in Table 5. Since the variances are close in each level between two distortion types, we only list the MOS here. For Dance Kiss, the MOS of scaling distortion is close to that of compression distortion. For Fox Bird, we observe a significant MOS drop in scaling distortion when the bit rate becomes low. Fox Bird is a bright video clip that contains stronger edges as a result of the cartoon content. The scaling distortion is more visible in a bright scene with strong edges. In contrast, Dance Kiss is a dark video clip in our selection. It has smoother textures. The scaling distortion is reduced by the dark scene and the smooth texture.

Since the MCL-V database contains a wide range of video contents, it can capture the characteristics of the human visual system better and allow researchers to develop better objective video quality assessment algorithms.

4.1. Performance comparison of objective VQA methods

Several full-reference (FR) IQA and VQA algorithms [57–64] are evaluated against the collected MOS of the MCL-V database and reported in this subsection. IQA methods can be extended to VQA methods by averaging frame-level quality scores. The IQA source codes of [57–59] are downloaded from [65]. Others are downloaded from respective authors' websites. The state-of-the-art VQA methods take both spatial and temporal artifacts into account. For example, the VADM method decouples two spatial distortion types; i.e. detail losses and additive impairments, and evaluate them separately. Furthermore, the motion information is adopted by VADM to measure the temporal masking effect. The ST-MAD method employs the spatio-temporal images to model the interaction between these two artifacts.

Three performance measure for these IQA and VQA methods are calculated and compared. They are: (1) the Pearson Correlation Coefficient (PCC) [25,31], (2) the Spearman rank-order correlation coefficient (SROCC) [25,31], and (3) the root mean squared error (RMSE) [25,31]. The PCC and SROCC are computed after nonlinear regression on the quality scores using the logistic function as recommended in [66]. Mathematically, we have

$$y = \beta_1 \cdot \left(0.5 - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}}\right) + \beta_4 \cdot x + \beta_5, \qquad (3)$$

where x is an objective quality score and $\beta_i,\ i = 1 \ldots 5$, are fitting parameters.

First, the performance of these quality metrics with respect to the compression distortion is shown in Table 6. We see that FSIM and VADM give the best performance among the test group for the compression distortion due to their good distortion models. They are close in PCC and RMSE while VADM provides a better SROCC measure. However, their PCC and RMSE values are still lower than 0.75, which allows room for further improvement.

Next, the performance of these quality metrics with respect to the scaling distortion is given in Table 7. First, we see that these metrics perform worse for the scaling distortion than the compression distortion. Second, VADM and FSIM are still the top two performers among the test group while VADM outperforms FSIM in all three scores.

Finally, we list the performance of all methods against the entire MCL-V database that contains both compression and scaling distortion types in Table 8. Furthermore, we list their performance against the LIVE database [23] for side-by-side comparison. The top three performers for the LIVE database are VADM, ST-MAD and T-MAD. Their PCC and SROCC scores are all above 0.80. In contrast, their performance degrades substantially in the MCL-V database, which indicates that MCL-V is a more challenging video quality database. This can be explained by that the source video in MCL-V is more diversified, and it is not easy to find an ideal metric to cover all of them.

Table 6
Performance comparison of objective quality metrics with respect to the compression distortion in MCL-V.

	PCC	SROCC	RMSE
PSNR	0.471	0.422	1.994
MSSIM [58]	0.617	0.609	1.779
SSIM [57]	0.650	0.633	1.718
VIF [59]	0.667	0.637	1.685
GMSD [62]	0.653	0.644	1.712
GSM [61]	0.715	0.713	1.580
FSIM [60]	0.770	0.775	1.441
S-MAD [64]	0.702	0.701	1.609
T-MAD [64]	0.625	0.623	1.763
ST-MAD [64]	0.657	0.663	1.702
VADM [63]	0.747	0.735	1.515

Table 7
Performance comparison of objective quality indices with respect to the scaling distortion in MCL-V.

	PCC	SROCC	RMSE
PSNR	0.463	0.493	1.881
MSSIM [58]	0.609	0.630	1.683
SSIM [57]	0.635	0.649	1.639
VIF [59]	0.636	0.661	1.637
GMSD [62]	0.634	0.662	1.642
GSM [61]	0.692	0.707	1.531
FSIM [60]	0.722	0.702	1.468
S-MAD [64]	0.659	0.624	1.594
T-MAD [64]	0.580	0.548	1.728
ST-MAD [64]	0.617	0.585	1.669
VADM [63]	0.728	0.741	1.469

Table 8
Performance comparison of objective quality indices with respect to both compression and scaling distortions in MCL-V.

	PCC		SROCC		RMSE	
Database	MCL-V	LIVE [23]	MCL-V	LIVE [23]	MCL-V	LIVE [23]
PSNR	0.472	0.549	0.464	0.523	1.956	9.176
MSSIM [58]	0.621	0.739	0.623	0.732	1.740	7.398
SSIM [57]	0.650	0.542	0.648	0.525	1.687	9.223
VIF [59]	0.660	0.570	0.655	0.557	1.666	9.019
GMSD [62]	0.650	0.737	0.661	0.726	1.686	8.414
GSM [61]	0.709	0.650	0.711	0.684	1.565	8.341
FSIM [60]	0.750	0.690	0.755	0.689	1.466	8.240
S-MAD [64]	0.681	0.737	0.670	0.721	1.624	7.669
T-MAD [64]	0.600	0.818	0.584	0.815	1.774	6.562
ST-MAD [64]	0.634	0.830	0.623	0.824	1.714	6.133
VADM [63]	0.742	0.844	0.752	0.835	1.489	5.945

5. Conclusion and future work

The construction of a new HD video quality assessment database, called MCL-V, was described in this work. MCL-V contains 12 source video clips and 96 distorted video clips with subjective assessment scores. The source video clips were selected from a large pool of public-domain HD video sequences with representative and diversified contents. Several existing IQA and VQA algorithms were evaluated against the MCL-V database. The database is publicly available at http://mcl.usc.edu/mcl-v-database/ for future research and development.

We attempted to analyze the relationship between video properties and the MOS values using 4 video sequences as examples in Section 4. A thorough analysis of the acquired MOS involves visual salience detection/tracking and a good understanding of the spatial/temporal masking effects. Although this is beyond the scope of our work, it is an interesting topic for further study. Furthermore, as shown in Section 4.1, there is no objective quality metric that has a PCC (or SROCC) value higher than 0.75 against the MCL-V database. The development of a better VQA method is also in need.

Acknowledgments

This project is funded by Netflix. The authors are grateful to Mr. David Ronca and Dr. Anne Aaron of Netflix for their valuable suggestions in various stages of this research. This work is supported by NSFC – China Grant No. 61401337, Fundamental Research Funds for the Central Universities, and China Scholarship Council. Computation for the work described in this paper was supported by the University of Southern California Center for High-Performance Computing and Communications (hpcc.usc.edu).

References

[1] D. Tice, How People Use Media: Over-the-Top TV 2013, Tech. rep., GfK Media (Aug 2013).

[2] S.E. Bird, The audience in everyday life: living in a media world, Routledge, 2013.

[3] H. Boujut, J. Benois-Pineau, O. Hadar, T. Ahmed, P. Bonnet, Weighted-mse based on saliency map for assessing video quality of h.264 video streams, in: Proc. SPIE, 2011, pp. 78670X-78670X-8.

[4] F. Boulos, W. Chen, B. Parrein, P. Le Callet, Region-of-interest intra prediction for H. 264/AVC error resilience, in: Image Processing (ICIP), 2009 16th IEEE International Conference on, IEEE, 2009, pp. 3109–3112.

[5] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, T. Ebrahimi, Subjective assessment of H. 264/AVC video sequences transmitted over a noisy channel, in: Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on, IEEE, 2009, pp. 204–209.

[6] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, T. Ebrahimi, A H.264/AVC video database for the evaluation of quality metrics, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 2430–2433.

[7] X. Feng, T. Liu, D. Yang, Y. Wang, Saliency based objective quality assessment of decoded video affected by packet losses, in: Image Processing (ICIP), 2008 15th IEEE International Conference on, IEEE, 2008, pp. 2560–2563.

[8] L. Goldmann, F. De Simone, T. Ebrahimi, A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video, in: Proc. SPIE, 2010, pp. 75260S–75260S–11.

[9] A. Khan, L. Sun, E. Ifeachor, J.O. Fajardo, F. Liberal, Impact of RLC losses on quality prediction for H.264 video over UMTS networks, in: Multimedia and Expo (ICME), 2010 IEEE International Conference on, IEEE, 2010, pp. 702–707.

[10] J.-S. Lee, F. De Simone, T. Ebrahimi, Subjective quality evaluation via paired comparison: application to scalable video coding, IEEE Trans. Multimedia 13 (5) (2011) 882–893.

[11] J.-S. Lee, F. De Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, T. Ebrahimi, Subjective evaluation of scalable video coding for content distribution, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 65–72.

[12] T. Liu, Y. Wang, J.M. Boyce, H. Yang, Z. Wu, A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts, IEEE J. Selected Top. Signal Process. 3 (2) (2009) 280–293.

[13] A. Moorthy, L.K. Choi, A. Bovik, G. de Veciana, Video quality assessment on mobile devices: Subjective, behavioral and objective studies, Select. Top. Signal Process., IEEE J. 6 (6) (2012) 652–671.

[14] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, Y. Wang, Modeling the impact of frame rate on perceptual quality of video, in: Image Processing (ICIP), 2008 15th IEEE International Conference on, 2008, pp. 689–692.

[15] Y.-F. Ou, Z. Ma, T. Liu, Y. Wang, Perceptual quality assessment of video considering both frame rate and quantization artifacts, IEEE Trans. Circuit. Syst. Video Technol. 21 (3) (2011) 286–298.

[16] Y.-F. Ou, Y. Zhou, Y. Wang, Perceptual quality of video with frame rate variation: A subjective study, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 2446–2449.

[17] S. Péchard, R. Pépion, P. Le Callet, et al., Suitable methodology in subjective video quality assessment: a resolution dependent paradigm, in: Proceedings of the Third International Workshop on Image Media Quality and its Applications, IMQA2008, 2008.

[18] Y. Pitrey, M. Barkowsky, P. Le Callet, R.Pépion, Evaluation of mpeg4-svc for qoe protection in the context of transmission errors, in: Proc. SPIE, 2010, pp. 77981C–77981C–9.

[19] Y. Pitrey, M. Barkowsky, P. Le Callet, R. Pepion, Subjective quality assessment of mpeg-4 scalable video coding in a mobile scenario, in: Visual Information Processing (EUVIP), 2010 2nd European Workshop on, IEEE, 2010, pp. 86–91.

[20] Y. Pitrey, M. Barkowsky, P. Le Callet, R. Pepion, Subjective Quality Evaluation of H. 264 High-Definition Video Coding versus Spatial Up-Scaling and Interlacing, QoE for Multimedia Content Sharing.

[21] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, P. Le Callet, Subjective quality of SVC-coded videos with different error-patterns concealed using spatial scalability, in: Visual Information Processing (EUVIP), 2011 3rd European Workshop on, IEEE, 2011, pp. 180–185.

[22] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, P. Le Callet, Aligning subjective tests using a low cost common set, QoE for Multimedia Content Sharing.

[23] K. Seshadrinathan, R. Soundararajan, A. Bovik, L. Cormack, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process. 19 (6) (2010) 1427–1441.

[24] N. Staelens, G. Van Wallendael, R. Van de Walle, F. De Turck, P. Demeester, High definition H. 264/AVC subjective video database for evaluating the influence of slice losses on quality perception, in: Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on, IEEE, 2013, pp. 130–135.

[25] Video Quality Experts Group (VQEG), Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I, Tech. rep. (2000).

[26] Video Quality Experts Group (VQEG), Report on the validation of video quality models for high definition video content, Tech. rep. (2010).

[27] J. You, T. Ebrahimi, A. Perkis, Modeling motion visual perception for video quality assessment, in: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, ACM, 2011, pp. 1293–1296.

[28] F. Zhang, S. Li, L. Ma, Y.C. Wong, K.N. Ngan, IVP subjective quality video database, <http://ivp.ee.cuhk.edu.hk/research/database/subjective/> (2011).

[29] T. Ebrahimi, Quality of multimedia experience: Past, present and future, in: Proceedings of the 17th ACM International Conference on Multimedia, MM '09, ACM, New York, NY, USA, 2009, pp. 3–4.

[30] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE J. Select. Top. Signal Process. 6 (6) (2012) 616–625.

[31] Video Quality Experts Group (VQEG), Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II, Tech. rep. 2003.

[32] ITU, Recommendation ITU-T P.910, Subjective video quality assessment methods for multimedia applications, International Telecommunication Union, Geneva, Switzerland 910.

[33] ITU, Recommendation ITU-R BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland.

[34] M.D. Brotherton, Q. Huynh-thu, D.S. Hands, K. Brunnstrom, Subjective multimedia quality assessment, IEICE Trans. Fundament. Electron., Commun. Comput. Sci. E89-A (11) (2006) 2920–2932.

[35] T. Tominaga, T. Hayashi, J. Okamoto, A. Takahashi, Performance comparisons of subjective quality assessment methods for mobile video, in: 2010 2nd International Workshop on Quality of Multimedia Experience, QoMEX 2010, IEEE, Trondheim, Norway, 2010, pp. 82–87.

[36] F. Bellard, M. Niedermayer, FFmpeg, <http://ffmpeg.org> (2012).

[37] J. Ohm, G.J. Sullivan, H. Schwarz, T.K. Tan, T. Wiegand, Comparison of the coding efficiency of video coding standards-including high efficiency video coding (hevc), IEEE Trans. Circuit. Syst. Video Technol. 22 (12) (2012) 1669–1684.

[38] C. Keimel, A. Redl, K. Diepold, The TUM high definition video datasets, in: Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on, 2012, pp. 97–102.

[39] CDVL, The Consumer Digital Video Library, <http://www.cdvl.org/> (2010).

[40] European Broadcasting Union (EBU), EBU HDTV Test Sequences, <http://tech.ebu.ch/> (2006).

[41] S. Goedegebure, A. Goralczyk, E. Valenza, N. Vegdahl, W. Reynish, B.V. Lommel, C. Barton, J. Morgenstern, T. Roosendaal, Big Buck Bunny, <http://www.bigbuckbunny.org/> (2008).

[42] L. Haglund, The SVT high definition multi format test set, Swedish Television Stockholm.

[43] L. Aimar, L. Merritt, E. Petit, M. Chen, J. Clay, M. Rullgrd, C. Heine, A. Izvorski, x264 – a free H.264/AVC encoder, <http://www.videolan.org/developers/x264.html> (2005).

[44] D.A. Silverstein, J.E. Farrell, Quantifying perceptual image quality, in: PICS, Vol. 98, 1998, pp. 242–246.

[45] D.A. Silverstein, J.E. Farrell, Efficient method for paired comparison, J. Electron. Imag. 10 (2) (2001) 394–398.

[46] J.-S. Lee, L. Goldmann, T. Ebrahimi, Paired comparison-based subjective quality assessment of stereoscopic images, Multimedia Tools Appl. 67 (1) (2013) 31–48.

[47] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, TID2008-a database for evaluation of full-reference visual quality assessment metrics, Adv. Modern Radioelectron. 10 (4) (2009) 30–45.

[48] J. Li, M. Barkowsky, P. Le Callet, Subjective assessment methodology for preference of experience in 3dtv, in: IVMSP Workshop, 2013 IEEE 11th, 2013, pp. 1–4.

[49] J. Li, M. Barkowsky, P. Le Callet, Analysis and improvement of a paired comparison method in the application of 3dtv subjective experiment, in: Image Processing (ICIP), 2012 19th IEEE International Conference on, 2012, pp. 629–632.

[50] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, W. Lin, Random partial paired comparison for subjective video quality assessment via hodgerank, in: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, ACM, New York, NY, USA, 2011, pp. 393–402.

[51] R.A. Bradley, Rank analysis of incomplete block designs: II. Additional tables for the method of paired comparisons, Biometrika 41 (3) (1954) 502–537.

[52] R.A. Bradley, M.E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (3) (1952) 324–345.

[53] K. Tsukida, M.R. Gupta, How to analyze paired comparison data, Tech. rep., DTIC Document (2011).

[54] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, TID2008-a database for evaluation of full-reference visual quality assessment metrics, Adv. Modern Radioelectron. 10 (4) (2009) 30–45.

[55] H. Gulliksen, A least squares solution for paired comparisons with incomplete data, Psychometrika 21 (2) (1956) 125–134.

[56] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction, Mathemat. Intell. 27 (2) (2005) 83–85.

[57] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[58] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, Conference Record of the Thirty-Seventh Asilomar

Conference on Signals, Systems and Computers, 2004, Vol. 2, IEEE, pp. 1398–1402.

[59] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[60] L. Zhang, D. Zhang, X. Mou, D. Zhang, Fsim: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.

[61] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Trans. Image Process. 21 (4) (2012) 1500–1512.

[62] W. Xue, L. Zhang, X. Mou, A. Bovik, Gradient magnitude similarity deviation: a highly efficient perceptual image quality index, IEEE Trans. Image Process. 23 (2) (2014) 684–695.

[63] S. Li, L. Ma, K.N. Ngan, Full-reference video quality assessment by decoupling detail losses and additive impairments, IEEE Trans. Circuit. Syst. Video Technol. 22 (7) (2012) 1100–1112.

[64] P. Vu, C. Vu, D. Chandler, A spatiotemporal most-apparent-distortion model for video quality assessment, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, 2011, pp. 2505–2508.

[65] M. Gaubatz, S. Hemami, MeTriX MuX visual quality assessment package, <http://foulard.ece.cornell.edu/gaubatz/metrix_mux> (2011).

[66] S. Chikkerur, V. Sundaram, M. Reisslein, L.J. Karam, Objective video quality assessment methods: a classification, review, and performance comparison, IEEE Trans. Broadcast. 57 (2) (2011) 165–182.