FULL REFERENCE VIDEO QUALITY ESTIMATION FOR VIDEOS WITH DIFFERENT SPATIAL RESOLUTIONS

A. Murat Demirtas[1], *Amy R. Reibman*[2], *and Hamid Jafarkhani*[1]

Center for Pervasive Communications and Computing, University of California, Irvine [1]
AT&T Labs - Research [2]

ABSTRACT

Full reference video quality estimators (QEs) either resize the input video or the reference video to compute the quality when these videos have different spatial resolutions. This resizing operation causes several limitations. Multiscale Image Quality Estimator (MIQE) [1] overcomes those limitations for images but it does not consider the temporal characteristics of video. In this work, we develop a video quality estimator that integrates MIQE with the motion information to estimate the quality. We also perform subjective tests to compare the proposed algorithm with the existing QEs. Test results show that the proposed algorithm outperforms other QEs.

Index Terms— video quality estimation, human visual system, motion, spatial resolution, subjective tests

1. INTRODUCTION

Multimedia communications have become ubiquitous as a result of progress in consumer products and social media abilities. Since the demand is increasing exponentially, the available wireless communication infrastructure may not be sufficient to provide a satisfactory service in the future. Therefore, we should develop methods to employ the resources efficiently. One of these methods is using perceptual coding approaches to minimize the required bitrate without causing any noticeable artifact. To achieve this, the quality of the visual content must be estimated with high accuracy.

Quality estimators (QEs) are used to measure the perceived quality of the visual content. QE approaches are evaluated in three categories. They are: Full-Reference (FR) QE, Reduced-Reference (RR) QE and No-Reference (NR) QE. To decide the type of the QE and design it accurately, we must take into account the requirements of the environment [2]. The display resolution of the end user is one of these important requirements if the network consists of devices with various spatial resolution constraints.

To date, FR video QEs have been designed to estimate quality when the reference and input videos have the same spatial resolutions. There are two straightforward methods to adapt these QEs to estimate the quality when the displayed frame has a different spatial resolution. The first method compares the low-resolution input video with the decimated reference video. The second method compares the interpolated input video with the reference video. Throughout this paper, we call these methods QE$_{down}$ and QE$_{up}$, respectively.

QE$_{down}$ and QE$_{up}$ have significant drawbacks [3]. The first drawback occurs because of interpolating the test or decimating the reference video to estimate the quality. These operations can cause over or under estimation of the input video's quality in QE$_{down}$ and QE$_{up}$ setups, respectively. The second drawback arises due to potential mismatch between filters which are used in creating the input video and computing the quality. Lastly, ignoring the effect of viewing distance to the quality is the third drawback. More details about these drawbacks and how they affect the quality computation can be found in [1]. While computing the video quality we should also consider the effect of motion to the quality and its influence on the aforementioned drawbacks.

The effect of spatial resolution on the quality has been examined in several works. In [4] and [5], the authors perform subjective tests to examine the *subjective* impact of jointly adjusting spatial resolution, temporal resolution, and quantization step-size. Cermak et al. [6] used the mean opinion scores (MOSs) obtained for QIF(176×144), CIF(352×288), VGA(640×480), and HD(1920×1200) resolutions at many bit rates. Cranley et al. [7] perform subjective tests using several combinations of resolution and quantization parameters to find the optimum parameter pairs. These studies provide valuable information to understand the effect of spatial resolution, but they do not provide an objective metric model that is used to quantify this effect. In [1], we propose Multiscale Image Quality Estimator (MIQE) to calculate the quality when the input and reference images have different resolutions. It handles the limitations of QE$_{down}$ and QE$_{up}$ approaches. However, using MIQE without considering the effect of motion will not provide an accurate video QE for videos with different spatial resolutions.

In this work, we propose a video quality estimator for input videos with different resolutions from the reference videos. It is designed to overcome the drawbacks of QE$_{down}$ and QE$_{up}$. It extends [1] to consider the video and in particular the effect of motion to the sensitivity of the human

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

vision system (HVS). We also develop a subjective test environment to ensure whether the proposed algorithm works and compare it with other video QEs. In Section 2, we describe the proposed quality estimator. In Section 3, we explain our subjective test environment and analyze the subjective test results. In Section 4, we conclude the paper.

2. MULTISCALE VIDEO QUALITY ESTIMATOR

The design of the proposed video quality estimator is based on MIQE [1] to handle the limitations of QE$_{down}$ and QE$_{up}$. However, it also takes into account the temporal characteristics of the video to the perceived quality. Hence, we develop the Multiscale Video Quality Estimator (MVQE) by incorporating the effect of motion information into the MIQE. We describe how we integrate motion into MIQE in detail in the following two subsections. First, we explain how the contrast sensitivity of HVS changes according to the spatial and temporal frequencies. Next, we use the spatiotemporal effects of HVS to construct our video QE.

2.1. The Spatiotemporal Effect of HVS on Video Quality

To design a video QE which takes into account the effect of motion and the viewing distance, we should employ the spatiotemporal contrast sensitivity function (STCSF) in the QE. STCSF shows the thresholds of HVS at different spatial and temporal frequency pairs to detect the changes in a visual content. We need proper frequency units to represent and compute the STCSF appropriately.

We use the angular frequency to represent spatial frequency components of two video frames that are viewed with different resolutions and/or different distances. The angular frequency is computed as follows [8]:

$$f(l) = \frac{\pi * d * n}{180 * h * 2 * 2^l} \tag{1}$$

In this expression, $f(l)$ denotes the angular (spatial) frequency in cycles per degree (cyc/deg); d, h, and n represent the distance of the viewer, height of the screen, and the number of pixels in the vertical direction, respectively. Lastly, l indicates the level of a subband decomposition. The temporal frequency is obtained by multiplying the magnitude of the velocity with the spatial frequency [9]. We use Daly's model [10] to calculate the frequency response of STCSF as follows:

$$\begin{aligned} &STCSF(f, v_R) = k.c_0.c_2.v_R.(c_1.2.\pi.f)^2.exp(-\frac{c_1.4.\pi.f}{f_{max}}) \\ &k = s_1 + s_2.|\log(\frac{c_2.v_R}{3})|^3 \\ &f_{max} = \frac{f_1}{c_2.v_R + 2} \end{aligned} \tag{2}$$

where $s_1 = 6$, $s_2 = 7.3$, $f_1 = 45.9$, $c_0 = 1.14$, $c_1 = 0.67$ and $c_2 = 1.92$. In this equation, f is the spatial frequency in cyc/deg and it is computed as in Eq. (1). The retinal velocity is denoted by v_R and measured in deg/sec. We compute v_R using the following equation:

$$v_R = v_T - v_E \tag{3}$$

In this expression, v_T is the velocity of the target object in deg/sec. It depends on the frame rate of the video and spatial frequency. It is computed as follows:

$$v_T(f) = m.FR/f \tag{4}$$

where, m is the estimated motion of the target object, FR is the frame rate of the video and f is the spatial frequency. On the other hand, v_E denotes the velocity of the eye. Daly's model takes into account the movement of the eye, and computes v_E as follows:

$$v_E = \min[(g_{sp}.v_T) + v_{MIN}, v_{MAX}] \tag{5}$$

where, $g_{sp} = 0.82$, $v_{MIN} = 0.15$ deg/sec, and $v_{MAX} = 80$ deg/sec. Using Eqs. (2)-(5), we can find the sensitivity of each spatial frequency and velocity pair.

While computing the quality of videos that have different resolutions, we should also take into account the effect of resizing on the computation of the spatial frequency of the visual content. As seen in Eq. (1), increasing the viewing distance or decimating by 2 doubles the spatial frequency of the visual content. Since temporal frequency depends on the motion and spatial frequency, resizing or distance change also indirectly affects the temporal frequency. In addition to the usage of the correct spatial resolution, accurate motion prediction is also necessary to find the sensitivity with high precision. To compute the STCSF value accurately, we need to use the methods which take into account the motion in different spatial frequency bands. We have selected the Hierarchical Block-Based Motion Estimation using wavelets (HBME).

HBME uses the correlation between the blocks in the consecutive two frames of the video. The first one is called the anchor frame and the second one is called the target frame. These two frames are decomposed into subbands using wavelet-based transforms. The motion is estimated for each subband using block-based ME. It is also assumed that there is a correlation in motion among subbands. Hence, the motion information which is found in the lower level is used as a starting point in the next level. The estimation starts with the lowest subband level and proceeds as follows.

First of all, the subband of the target frame and anchor frames are divided into non-overlapping 4×4 blocks. Second, a destination block is picked from the target frame's subband. This block is called the destination block. Third, a search range is defined in the anchor frame's subband for the corresponding block. The center of this range collocates with the center of the destination block. Fourth, all the candidate blocks within the search are compared with the destination block. The comparison is performed using the mean absolute difference (l_1 norm). The index of the block which has the minimum difference determines the motion vector for the chosen destination block. These operations are repeated until all the motion vectors are computed. After motion vectors of all blocks are found in the lowest subband level, they are scaled and used as starting locations for the motion estimation in the upper layers.

2.2. Quality Calculation

In this section, we describe how we compute quality using MVQE. The first steps are identical to MIQE [1]: convert pixel values to physical luminance, compute a wavelet decomposition on reference and input videos, and represent blocks from the input in terms of the reference using a Gaussian Mixture Model. Next, we construct the QE by appropriately weighting the mutual information of each subband block. The calculation of mutual information for each subband block is similar to what we did for images in [1] but the the weights of subband blocks are computed differently. In MIQE, weighting coefficients only depend on the subband level. Therefore, all subband blocks in the same subband will have the same weights. On the other hand, in MVQE, the weight of each subband block is computed separately, because the sensitivity function depends on the motion and the motion may vary according to the position of the subband block in the video frame.

To calculate the magnitude of weighting coefficients, we take into account two factors. First, the effect of subband size should be compensated by scaling the mutual information of each subband block with 2^{2l}. Second, HVS has a different weight for each subband block. As expressed in Eq. (2), STCSF depends on the motion and the spatial frequency. Since the motion information is local and it can be different for different spatial frequencies, the value of the contrast sensitivity can change for each block. Hence, the scaling coefficient of each subband block is computed using the following expression:

$$G_{l,o,j} = STCSF(l,o,j) * 2^{2l} \tag{6}$$

where l and o, denote the subband level and the subband orientation, respectively. The term j stands for the block index at the subband (l, o). The value of the estimated video quality is found by scaling the similarity of each subband block with the corresponding weight. It is computed as follows:

$$MVQE = \frac{\sum_{l,o,j \,\in\, subband\ blocks} G_{l,o,j} * I_{I,l,o,j}}{\sum_{l,o,j \,\in\, subband\ blocks} G_{l,o,j} * I_{R,l,o,j}} \tag{7}$$

$I_{I,l,o,j}$ and $I_{R,l,o,j}$ represent the mutual information values of the input and the reference signals for the subband block (l,o,j), as calculated in [1]. The final quality value is between 0 and 1, and increases as the quality improves. To check the validity of the proposed algorithm, it is necessary to compare the estimated quality values with the viewers' opinion scores.

3. SUBJECTIVE TESTS

We performed subjective tests to obtain the viewers' mean opinion scores using videos that have different spatial resolutions. The test results are used to ensure that the proposed algorithm works properly. They also provide a benchmark to compare our algorithm with existing approaches. In the following subsections, we describe how we create the test set, perform subjective tests, and analyze results.

3.1. Test Set Creation

We use in total 4 reference videos with different spatial and temporal complexities. They are chosen from a public database in [11] where the original sources are referenced. The name of the videos are *Soccer*, *Mobcal*, *Tree* and *Park*. *Mobcal*, *Tree* and *Park* have a 1280×720 spatial resolution; their frame rate is 50 fps. On the other hand, the spatial resolution of *Soccer* is 704×576 and its frame rate is 60 fps. The videos have different spatial and temporal characteristics.

We have created 9 test videos have been created for each reference video. Out of these 9 test videos, 5 are high-resolution and are created by compressing a reference video. The remaining 4 videos are low-resolution test videos. These videos are obtained by decimating the reference video at a rate of 0.5 and compressing the decimated video. We compress the full-size reference video and the decimated reference video using H.264 AVC codec. JM 18.1 [12] reference software is used to implement H.264 AVC codec. We employ the Non-Normative Filter as a low pass filter during decimation. It is a Sine-windowed Sinc-function and is formulated in [13].

Quantization levels for the test videos should cover a range from excellent quality to very bad quality. Moreover, we should also observe the effect of bitrate on the quality of high and low resolution test videos at different quality regions. To fulfill these two requirements, QP values of the low resolution test videos are determined according to the QP values of the high resolution test videos. We use the following notations to describe the relationships between the bitrate and the quantization values of high-resolution and low-resolution test videos. HR and LR denote high-resolution and low-resolution test videos, respectively. The quantization level index is denoted by i. Hence, QP_i^{HR} represents the QP value of high-resolution video at the i^{th} index, and BR_i^{HR} denotes the bitrate of this video. The QP value increases (quality decreases) as i increases.

Based on this notation, we choose the bitrates of the test video as follows. First of all, we assume that $QP_1^{HR} = 28$ and $QP_1^{LR} = 28$ for all reference video types. If i is even $(i \in 2, 4)$, then BR_i^{HR} will be found by averaging BR_{i-1}^{HR}

Table 1. Mean Opinion Scores and 95 % CIs

	H1	L1	H2	L2	H3	L3	H4	L4	H5
Mobcal	88.9 ± 2.4	74.2 ± 4.9	83.5 ± 3.2	70.4 ± 4.6	63.7 ± 5.1	48.2 ± 5.3	43.0 ± 7.0	30.8 ± 6.4	31.7 ± 7.6
Soccer	90.5 ± 3.1	73.8 ± 6.3	86.7 ± 3.0	64.2 ± 6.6	70.7 ± 4.4	53.0 ± 8.1	58.2 ± 6.1	42.5 ± 6.6	36.9 ± 6.4
Tree	88.8 ± 2.6	71.7 ± 5.1	82.4 ± 3.5	62.0 ± 5.9	60.8 ± 6.4	42.7 ± 7.3	42.7 ± 7.9	28.8 ± 6.7	20.8 ± 6.7
Park	89.1 ± 2.7	72.8 ± 5.1	85.0 ± 3.3	66.2 ± 6.3	65.2 ± 5.6	43.0 ± 7.0	48.6 ± 7.3	27.6 ± 7.5	24.6 ± 7.8

and BR_{i-1}^{LR}. QP_i^{HR} and QP_i^{LR} are equal to the QP value of the high-resolution test video which has the closest bitrate to the BR_i^{HR}. On the other hand, if i is odd ($i \in 3, 5$), we set BR_i^{HR} to BR_{i-2}^{LR} and the QP value which satisfies this requirement is assigned to QP_i^{HR}.

To have reliable quality estimates without compromising the accuracy of ranking, the viewers should be able to playback the videos anytime during the test. The method that we use to achieve this goal is the Subjective Assessment Methodology for Video Quality (SAMVIQ) [14]. This method allows the viewer to select and watch the videos as many times as they want using the index of the video. Viewer can also update the scores based on other videos. During these updates, the viewer implicitly ranks the videos.

3.2. Subjective Test Implementation

The test contained a total of 40 videos, and 25 people joined the tests. Each test session took approximately 40 minutes. The viewers were either graduate or undergraduate students. They had clear vision and they were non-experts. The test session consisted of 4 groups. In each group, a reference video and its corresponding test videos were displayed to the viewers. There were 4 low resolution and 5 high resolution test videos for each group. The reference video could be selected by pressing 0. However, the indices of the test videos changed randomly for each test session and each test group to prevent bias. The order of test groups was also random.

We asked the viewers to give a score between 0 and 100 for each test video. The score of the reference video was 100. The range between 0 and 100 was divided into 5 parts as described in Double Stimulus Impairment Scale : Bad (0-20), Poor (21-40), Fair (41-60), Good (61-80), Excellent (81-100). We informed the viewers about the meaning of each range before the test. Each video took 10 seconds. Viewers could enter the scores before or after 10 seconds. The distance between the viewers and the screen was 6 times the height of the 640×360 image.

3.3. Analysis of Subjective Test Results

In this subsection, we compare the performance of our proposed video QE to 4 approaches using the results of the subjective test. We begin by evaluating the statistical characteristics of the subjective test. Then, we examine the similarity between subjective test scores and QE scores using QE_{down} setup.

Table 2. Correlation Metric Scores of QEs

	Pearson	SRCC	KRCC	FCR	FCS
PSNR	0.811	0.836	0.648	0.788	0.561
VQM	0.930	0.927	0.756	0.850	0.680
MOVIE	0.882	0.938	0.803	0.875	0.696
MIQE	0.940	0.952	0.819	0.925	0.784
MVQE	**0.969**	**0.973**	**0.883**	**0.975**	**0.876**

To assess the reliability of subjective test scores, we compute the 95% confidence interval (CI) of subjective test scores. According to these computations, when the QP level is high the variance of the MOS is also high. MOSs and CI values of each test video is given in Table 1.

Next, we calculate the similarity between MOSs and QEs using the following popular correlation metrics: Pearson, Spearman Rank Correlation Coefficient (SRCC), and Kendall Rank Correlation Coefficient (KRCC). These computations provide both the ranking and the scoring similarity. We also compute the Fraction of Correct Ranking (FCR) and Fraction of Correct Similarity (FCS) [1] metrics. Pearson, SRCC and KRCC consider the relationship between all test video scores, but FCR and FCS only examine the relationship between HR and LR test videos. We use QE_{down} setup to compute the QE values. During QE_{down}, we used the Non-Normative filter as the decimation filter. Compared QE methods are Video Quality Metric (VQM) [15], Motion-based Video Integrity Evaluation (MOVIE) [16], MIQE [1], and PSNR.

Table 2 shows the correlation scores between QEs and MOSs for all videos. According to the table, the PSNR has the lowest performance. MVQE performs better than other QEs. MIQE's correlation scores are less than those of MVQEs. Hence, integrating motion information has improved the QE estimation. FCR and FCS scores of the proposed approach is also higher than others.

4. CONCLUSIONS

Computing the quality of a video with lower spatial resolution compared to the reference video is a challenging task. In this paper, we have proposed an algorithm to solve this problem. We have developed this algorithm by incorporating the effect of motion to the our previously proposed method for images, i.e. MIQE. We have also performed subjective tests to measure the performance of the proposed approach. According to the test results, the proposed algorithm outperforms other QEs.

5. REFERENCES

[1] A. M. Demirtas, A. R. Reibman, and H. Jafarkhani, "Full-reference quality estimation for images with different spatial resolutions," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2069–2080, 2014.

[2] J. G. Apostolopoulos and A. R. Reibman, "The challenge of estimating video quality in video communication applications [in the spotlight]," *IEEE Signal Process. Mag.*, vol. 29, no. 2, pp. 160–, 2012.

[3] A. M. Demirtas, H. Jafarkhani, and A. R. Reibman, "Quality estimation for images and video with different spatial resolutions," in *Human Vision and Electronic Imaging XVII*, Feb. 2012, vol. 8291.

[4] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Toward optimal rate control: A study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," in *SPIE Visual Communications and Image Processing*, 2003, vol. 5150, pp. 198–209.

[5] J. S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[6] G. Cermak, M. Pinson, and S. Wolf, "The relationship among video quality, screen resolution, and bit rate," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 258–262, 2011.

[7] N. Cranley, P. Perry, and L. Murphy, "Optimum adaptation trajectories for streamed multimedia," *Multimedia Systems*, vol. 10, no. 5, pp. 392–401, August 2005.

[8] K.N.Ngan, K.S. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1743–1750, 1989.

[9] A. B. Watson and A. J. Ahumada, "Model of human visual-motion sensing," *Journal of The Optical Society of America*, vol. 2, no. 2, pp. 322–341, 1985.

[10] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Human Vision and Electronic Imaging III*, July 1998.

[11] "Xiph.org video test media," `http://media.xiph.org/video/derf/`.

[12] JVT, "Reference software version jm18.1," `http://iphome.hhi.de/suehring/tml/download/oldjm/jm18.1.zip`.

[13] S. Sun and J. Reichel, "AHG report on spatial scalability resampling," in *ISO/IEC JTC1/SC29/WG11, Doc JVT-R006*, 14 – 20 Jan. 2006.

[14] ITU, "Methodology for the subjective assessment of video quality in multimedia applications," Recommendation BT.1788, International Telecommunication Union, Geneva, 2007.

[15] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[16] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Broadcasting*, vol. 19, no. 2, pp. 335–350, 2010.