Video Quality Assessment by Decoupling Distortions on Primary Visual Information

Yang Li[1], Xu Wang[1], Feng Li[1], Qingrui Guo[1], Qiang Fan[2], Qiwei Peng[2], Wang Luo[2(✉)], Min Feng[2], Yuan Xia[2], and Shaowei Liu[2]

[1] State Grid Xinjiang Electric Power Science Research Institute, Nanjing 210003, China
[2] NARI Group Corporation (State Grid Electric Power Research Institute), Nanjing 210003, China
luowang@sgepri.sgcc.com.cn

Abstract. Video quality assessment (VQA) aims to evaluate the video quality consistently with the human perception. In most of existing VQA metrics, additive noises and losses of primary visual information (PVI) are decoupled and evaluated separately for quality assessment. However, PVI losses always include different types of distortions such that PVI distortions are not evaluated well enough. In this paper, a novel full-reference video quality metric is developed by decoupling PVI distortions into two classes: compression distortions and transmission distortions. First, video denoising method is adopted to decompose an input video into two portions, the portion of additive noises and the PVI portion. Then, maximal distortion regions searching (MDRS) algorithm is designed to decompose PVI losses into transmission distortions and compression distortions. Finally, the three distortions are evaluated separately and combined to compute the overall quality score. Experimental results on LIVE database show the effectiveness of the proposed VQA metric.

Keywords: Video quality assessment · Human visual system
Transmission distortion · Decoupling distortion

1 Introduction

Full Reference (FR) video quality assessment (VQA) metrics refer to algorithms that evaluate qualities of distorted videos with available reference videos. The goal is to evaluate the quality consistently well with human visual system (HVS). Signal-to-noise ratio (SNR) and peak SNR (PSNR) are the most widely used FR quality metrics. These indices are simple to calculate and convenient to be adopted. But they show poor consistency with subjective evaluations [1, 2].

Recently, the area of FR metrics has attracted a lot of attention [3–8]. Structure similarity index (SSIM) [3] was presented as a metric using structural information. To account for human perception of motion information, proper temporal weighting schemes [4, 5] were proposed based on SSIM. Temporal distortions were also considered in VQA metrics during the recent years. A motion-based video integrity

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
M. Huang et al. (Eds.): WICON 2016, LNICST 214, pp. 299–307, 2018.
https://doi.org/10.1007/978-3-319-72998-5_32

evaluation (MOVIE) index [6] was proposed to define the temporal distortion as the differences between the filter responses along computed motion trajectories. Spatio-temporal structural information was designed to evaluate the video perceptual quality [7]. Both the spatial edge features and temporal motion characteristics [9, 10] were accounted for with the structural features in the localized space-time regions. Different types of distortions cause different degradations. A decoupling based metric [8] were proposed by decomposing distortions into additive noises and distortions on primary visual information (PVI). The overall score was computed by combining evaluations of the two portions.

PVI distortions can be mainly classified into two types: compression distortions (e.g., Ringing, Blocking artifacts) and transmission distortions (e.g., Packet loss). However, these two types of distortions are with different characteristics. Compression distortions are content-dependent distortions. Structural similarity based metrics can perform well on this type of distortion. Transmission distortions which are introduced by packet loss are content-independent distortions. Large and distinct distortion regions randomly appear in video frames. Structural similarity based metrics, such as gradient similarity based metrics [7] can not represent transmission distortions accurately enough, especially when they occur in the original flat regions.

In this letter, a novel video quality metric is developed by decoupling PVI distortions. Video distortions are firstly decomposed into additive noises and PVI distortions using denoising method. Then, PVI distortions are classified into two typical classes: compression distortions and transmission distortions. After evaluating each type of distortions with rational metric, we combine the three evaluated results to compute the overall quality score. Experimental results on LIVE database show the effectiveness of the proposed VQA metric.

2 Proposed Method

We first give a brief overview of our approach before going into detail in subsequent sections. The flowchart of the proposed model is shown in Fig. 1. The reference video (V_r) and the test video (V_t) are firstly decomposed into additive noises (A_r and A_t) and

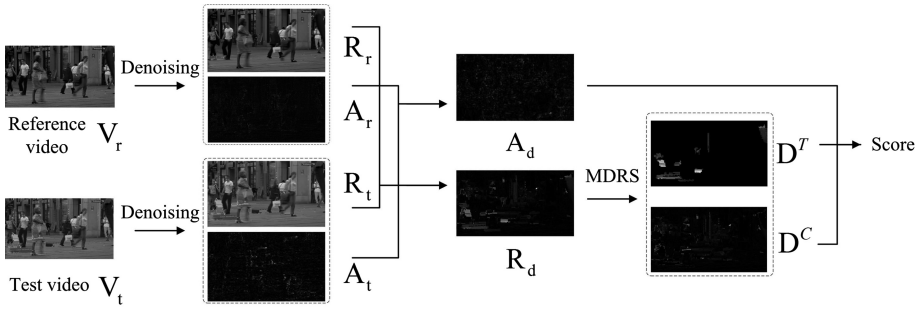

Fig. 1. Flowchart of the proposed model. V_r (V_t) is the reference (test) video, A_r (A_t) and R_r (R_t) are the additive noises and reconstructed portions of V_r (V_t), respectively. D^T and D^C are the portions of transmission distortions and compression distortions, respectively.

PVI portions (R_r and R_t) by using video denoising method. Then, PVI distortions (R_d, the difference between R_r and R_t) are classified into compression distortions (D^C) and transmission distortions (D^T), respectively. Different metrics are adopted to measure the degradations of each type, respectively. Finally, a combining scheme is used to generate the overall perceptual quality score of the test video.

2.1 Decoupling Distortions

Algorithm 1 : Maximal Distortion Regions Search (MDRS)

Input: The distortion of denoised video D.
Initialization: region number: K, difference threshold: T, window size: ws, area threshold: S

$OutIndex = cell(1, K)$;
$[W, H] = size(R_d)$; $M = zeros(W, H)$;
for $r = 1 : 4 : W$ **do**
  for $c = 1 : 4 : H$ **do**
    if $(length(find(R_d(r : r + 3, c : c + 3) > T)) == 16)$ **then**
      $M(r : r + 3, c : c + 3) = 1$;
    end if
  end for
end for
for $i = 1 \rightarrow K$ **do**
  $Idx = find\_max\_conn\_regions(M)$; $M(Idx) = 0$;
  if $length(Idx) > S$ **then**
    $OutIndex(i) = Idx$;
  end if
end for
Output: $OutIndex$;

The reference video and test video are firstly decomposed into additive noise portions and PVI portions via a video denoising method, called VBM3D [11]. A sparse 3D transform-domain collaborative filtering is used in VBM3D. This method performs good restoration on videos with additive noises. The differences between the additive noise portions of the reference video and test video are used to evaluate the additive distortions. This type of distortion may cause uncomfortable sensation. The differences between the PVI portions of the reference video and test video are called PVI distortions.

In order to represent the degradations more accurately, PVI distortions are decomposed into compression distortions and transmission distortions using a method, called Maximal distortion regions searching (MDRS). Compared to ringings and blocky artifacts which were introduced by compression, transmission errors always generate large areas of distortions. Furthermore, the strengths of these distortions are always large in local regions. That is, the regions with large area and distinct distortions tend to be with transmission errors. The goal of MDRS is to detect k maximal regions with transmission distortions, such that $k < K$, where K is the preset maximal number of regions. The PVI distortions R_d are the inputs. The outputs are the locations of the k

regions. Details of MDRS are shown in Algorithm 1. (1) To achieve the constraint that distortions in local regions are large, the PVI distortion (R_d) is split into 4 × 4 non-overlapping blocks. For each block, if all the absolute values of the differences are larger than the preset threshold T, flags in the corresponding block of M are set to 1. Otherwise, flags are set to 0. (2) The k maximal regions will be searched out by finding the maximal 1−connected regions in M. Areas of regions that are bigger than the preset threshold S are considered to be regions with transmission distortions. The other regions are with compression distortions. In this letter, K, S, and T are set to 8, 32, and 12, respectively, from our empirical study.

2.2 Additive Distortion Evaluation

Since MSE presents a good match with additive noises [8], MSE is adopted to evaluate the degradations of the additive noises as follows

$$S_A = S(A_r, A_t) = 1 - \log_{10}(1 + MSE(A_r, A_t))/\log_{10}(255^2), \tag{1}$$

where A_r and A_t are the additive portions of the reference video and test video, respectively; $MSE(A_r, A_t)$ is the mean squared error between A_r and A_t.

2.3 Transmission Distortion Evaluation

Transmission distortions are always introduced by packet losses. They are susceptible to the strength and area of the distortion regions. Therefore, transmission distortions are evaluated as follows

$$S_T(D^T) = \frac{C_1 - log_{10}\left(S \cdot T + \sum_{i=1}^{k} L_i \cdot S_i^2\right)}{C_1 - log_{10}(S \cdot T)}, \tag{2}$$

where D^T is the transmission distortion detected by MDRS method; Li, (i = 1, 2..k) is the mean absolute values of the i^{th} maximal distortion region; Si, (i = 1, 2..k) is the area of the i^{th} maximal distortion region; k is the region number; $C_1 = log_{10}(255 \cdot W^2 \cdot H^2)$, W and H are the image width and height, respectively; S and T are the preset area and difference thresholds used in MDRS. Similar with (1), the denominator is a normalization factor. Adding $S \cdot T$ is to avoid big leap occuring between the scores with and without transmission errors.

2.4 Compression Distortion Evaluation

Compression distortions mainly include degradations such as blurrings, blocking artifacts, and ringings. These degradations can be represented well by structural similarity based metrics, such as SSIM, edge gradient similarities. In this work, gradient similarities in spatial and temporal domain are computed to evaluate the compression

distortions. For each pixel, spatio-temporal gradient vector g = (g_x, g_y, g_t) is computed with Sobel filters along x, y and t directions, respectively. The Sobel Kernel for t direction is a 3 × 3 × 3 matrix [7]. To balance the effect of temporal and spatial gradients, they are divided by the sum of positive filter coefficients, respectively.

Since human attention mainly allocated to the salient regions, salient pixels are selected to evaluate the degradations of compression distortions. A pixel is considered to be a salient pixel if its spatio-temporal gradient magnitude is above the certain threshold ζ in either original video or distorted video [7]. The similarities between pixels can be computed as

$$S_p(x_r, x_t) = \left(\frac{2\|\mathrm{g}^r\|\|\mathrm{g}^t\| + C_2}{(\mathrm{g}^r)^2 + (\mathrm{g}^t)^2 + C_2} \right)^{\alpha} \cdot \left(\frac{\mathrm{g}^r \cdot \mathrm{g}^t + C_2}{\|\mathrm{g}^r\|\|\mathrm{g}^t\| + C_2} \right)^{\beta}, \tag{3}$$

where C_2 is the small constant to avoid the denominator being zero and is set as $C_2 = 0.03 \times 255^2$ [3]; g^r and g^t are the spatio-temporal gradient vectors of pixels x_r and x_t in the reference and test frames, respectively. In (3), the first term represents the similarity of magnitudes between g^r and g^t; the second term represents the similarity of directions between g^r and g^t. α and β are the relative importance of the two terms. A big value of α highlights the importance of vector magnitude. In this letter, α is set to 2 and β is set to 1.

Furthermore, HVS is highly sensitive to blocking artifacts. To measure the degradation of blocky artifacts, spatial gradient $\mathrm{g}^b = \left(g_x^b, g_y^b\right)$ similarities of down-sampled images are evaluated. The reference frame and test frame are down-sampled with rate 8 in both the vertical and horizontal directions. Blocking artifacts are evaluated as

$$S_b(b_r, b_t) = \left(\frac{2\left\|\mathrm{g}^{br}\right\|\left\|\mathrm{g}^{bt}\right\| + C_2}{\left(\mathrm{g}^{br}\right)^2 + \left(\mathrm{g}^{bt}\right)^2 + C_2} \right)^{\alpha} \cdot \left(\frac{\mathrm{g}^{br} \cdot \mathrm{g}^{bt} + C_2}{\left\|\mathrm{g}^{br}\right\|\left\|\mathrm{g}^{bt}\right\| + C_2} \right)^{\beta}, \tag{4}$$

where b_r and b_t are the pixels in the downsampled frames, i.e., mean values of 8 × 8 non-overlapped blocks of the reference and test frames, respectively. g^{br} and g^{bt} are spatial gradient vectors of b_r and b_t, respectively. α and β are set to the same values as in (4).

Some image regions have no apparent edge but are still with structural characteristics. Structural similarity [3] is adopted to evaluate the degradations on spatial structural information as

$$S_s(x_r, x_t) = \frac{2\sigma_{x_r, x_t} + C_2}{\sigma_{x_r}^2 \sigma_{x_t}^2 + C_2}, \tag{5}$$

where $S(x_r, x_t)$ is the structural similarity between blocks centered at pixels x_r and x_t. The block size is set to 11 × 11.

Combine pixel similarities, blocking artifacts, and structural similarities, compression distortion are deduced as

$$S_C = \underset{x_r \in Z}{\text{Avg}}\left(\sum \left(S_p(x_r, x_t) S_s(x_r, x_t) S_b(b_r, b_t)\right)\right). \tag{6}$$

where b_r (b_t) is the corresponding down-sampled value of the block in which pixel x_r (x_t) located; Z is the set of salient pixels which are degraded with compression distortions; Avg(·) is to calculate the average similarity of Z.

2.5 Overall Perceptual Quality

Distortion of different types will co-determine the perceptual quality of each frame. To evaluate the distortions on PVI, (2) and (6) are used to compute the perceptual quality score. Furthermore, the weights of the two evaluation parts are related to the noise energy level of the two portions. The similarities of additive distortions can reflect that of the compression distortions to some extent. Therefore, (1) are used to adjust the relative importance of the two portions. Finally, we combine the evaluation of the three portions, (1), (2), and (6). For each frame, the similarity can be computed as

$$S_F(V_r, V_t) = S_T^{S_A} \cdot S_C^{1-S_A} \tag{7}$$

Finally, all of the frame scores are averaged to give a final video quality index.

3 Experimental Results

In this section, the effectiveness of the proposed perceptual VQA metric is demonstrated. The LIVE subjective quality video database [12] is used to evaluate the performance of the proposed VQA metric. The LIVE database consists of 10 768 × 432p reference videos and 150 distortion videos. Subjective scores (DMOS) were recorded for all test sequences. The types of distortion comprised of MPEG-2 compression, H.264 compression, and simulated transmission of H.264 compressed bit-streams through error-prone IP networks and error-prone wireless networks. To detect the salient pixels, ζ is set to 300 through exhaustive experiments. For comparison, results with state-of-the-art VQA metrics are reported. The comparison metrics include PSNR, SW-SSIM [4], MC-SSIM [5], MOVIE [6], STSI [7], VQM [13], and Picture Quality Analyzer. Parts of the results are quoted from [7]. Pearson correlation coefficient (PCC) and Spearman rank order correlation coefficient (SROCC) are used as performance indicators.

The mapping function chosen for regression for each of the metrics is

$$f(x) = \frac{\beta_1 - \beta_2}{1 + exp\left(-\frac{x-\beta_3}{\beta_4}\right)} + \beta_2. \tag{8}$$

where $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ are the fitting parameters.

Table 1 shows the PCC and SROCC on the LIVE video quality database. It can be seen that the proposed metric significantly outperforms other metrics according to the two indicators. The PSNR performs especially poorly on this database. It shows that the conventional pixel based models are incapable to represent perceptual video quality.

Table 1. Performance comparison on the live database

Methods	Pearson CC	Spearman CC
VQM	0.702	0.723
MOVIE	0.786	0.810
STSI	0.779	0.778
SW-SSIM	0.585	0.596
MC-SSIM	0.679	0.698
PSNR	0.368	0.404
PQR (by PQA500)	0.695	0.712
DMOS (by PQA500)	0.695	0.711
Proposed	0.816	0.809

Table 2. Spearman CC Scores of VQA metrics on each kind of distortion in live database

Methods	Wireless	IP	H.264	MPEG2
PSNR	0.4675	0.4108	0.4385	0.3856
VQM	0.7325	0.6480	0.6459	0.7860
STSI	0.7544	0.8072	0.8298	0.6624
SW-SSIM	0.5867	0.5587	0.7206	0.6270
PQR (by PQA500)	0.6464	0.7300	0.7455	0.6456
DMOS (by PQA500)	0.6426	0.7295	0.7427	0.6445
Proposed	0.7786	0.8069	0.8792	0.7023

In Table 1, it also can be seen that the proposed VQA metric performs significantly better than the SSIM based metrics such as SW-SSIM (PCC increment: 0.23), and MC-SSIM (PCC increment: 0.14). The reason is that the proposed decoupling based method can detect different types of distortions and evaluate each type with the rational measure. Furthermore, the proposed method performs better than the spatio-temporal gradient similarity based method, STSI (Pearson CC increases by 0.04). This can be attributed to that gradient similarity based method cannot represent transmission errors accurately enough. Since transmission distortions are always flat regions without apparent edges, gradient similarity based methods cannot detect these distortions well enough, especially when transmission distortions occur in the original flat regions.

Table 2 shows the PCC on the four kinds of distortions in LIVE database. It demonstrates that the proposed metric performs the best on three kinds of distortions (Wireless, IP, and H.264). For MPEG2 distortion, even though it is not better than VQM metric, it performs significantly better than all the other metrics (SROCC

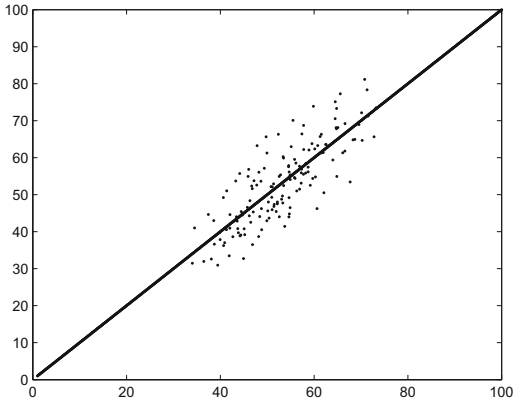

Fig. 2. Comparison between scatter plots of the proposed VQA.

increment: at least 0.04). The proposed metric is rather robust to various types of video distortions.

Figure 2 shows the scatter plots of the DMOS against the objective score by the proposed VQA metric on the LIVE database. It can be seen that the proposed metric performs well on videos from low quality to high quality.

4 Conclusion

In this letter, a VQA metric by decoupling PVI distortions has been proposed. Besides decoupling videos into additive noises and PVI, PVI distortions are decomposed into compression distortions and transmission distortions in order to evaluate PVI distortions more accurately. Considering the different properties of the decomposed portions, we separately evaluate their quality degradations with rational metrics. Experimental results demonstrate the effectiveness of the proposed metric.

References

1. Wang, Z., Bovik, A.: A universal image quality index. IEEE Signal Process. Lett. **9**(3), 81–84 (2002)
2. Zhang, X., Feng, X., Wang, W., Xue, W.: Edge strength similarity for image quality assessment. IEEE Signal Process. Lett. **20**(4), 319–322 (2013)
3. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
4. Wang, Z., Li, Q.: Video quality assessment using a statistical model of human visual speed perception. J. Opt. Soc. Am. A **24**(12), B61–B69 (2007)
5. Moorthy, A., Bovik, A.: Efficient video quality assessment along temporal trajectories. IEEE Trans. Circuits Syst. Video Technol. **20**(11), 1653–1658 (2010)
6. Seshadrinathan, K., Bovik, A.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans. Image Process. **19**(2), 335–350 (2010)

7. Wang, Y., Jiang, T., Ma, S., Gao, W.: Novel spatio-temporal structural information based video quality metric. IEEE Trans. Circuits Syst. Video Technol. **22**(7), 989–998 (2012)
8. Wu, J., Lin, W., Shi, G., Liu, A.: Perceptual quality metric with internal generative mechanism. IEEE Trans. Image Process. **22**(1), 43–54 (2013)
9. Xiong, J., Li, H., Wu, Q., Meng, F.: A fast HEVC inter CU selection method based on pyramid motion divergence. IEEE Trans. Multimedia **16**(2), 559–564 (2014)
10. Xiong, J., Li, H., Meng, F., Zhu, S., Wu, Q., Zeng, B.: MRF-based fast HEVC inter CU decision with the variance of absolute differences. IEEE Trans. Multimedia **16**(8), 2141–2153 (2014)
11. Dabov, K., Foi, A., Egiazarian, K.: Video denoising by sparse 3D transform-domain collaborative filtering. In: Proceedings of European Signal Processing Conference, EUSIPCO 2007, Poznan, Poland, September 2007
12. Seshadrinathan, K., Soundararajan, R., Bovik, A., Cormack, L.: Study of subjective and objective quality assessment of video. IEEE Trans. Image Process. **19**(6), 1427–1441 (2010)
13. Pinson, M., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Trans. Broadcasting **50**(3), 312–322 (2004)