VIDEO QUALITY ASSESSMENT BASED ON ADAPTIVE BLOCK-SIZE TRANSFORM JUST-NOTICEABLE DIFFERENCE MODEL

Lin Ma, Fan Zhang, Songnan Li, King N. Ngan

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
Email: {lma, fzhang, snli, knngan}@ee.cuhk.edu.hk

ABSTRACT

In this paper, we propose a full reference Video Quality Assessment (VQA) algorithm based on the Adaptive Block-size Transform Just-Noticeable Difference (ABT-JND) model. Firstly, ABT-JND is introduced for its efficiency of modeling the Human Vision System (HVS) characteristics. Based on the ABT-JND model, the full reference VQA is developed, by capturing HVS responses of spatio-temporal distortions over different block-size transforms. Experimental results have demonstrated that the proposed VQA outperforms other VQA methods, while slightly poorer than MOVIE. However, it maintains a very simple formulation. Since the proposed VQA performs on transform domain, it could be easily applied on many related applications, such as video compression, watermarking, and so on.

Index Terms— Video Quality Assessment (VQA), Adaptive Block-size Transform (ABT), Just-Noticeable Difference (JND), Human Vision System (HVS)

1. INTRODUCTION

Video quality control plays a very important role in many networked video applications, such as video compression, video on demand, digital television, video teleconferencing, and streaming video over the Internet, etc. Humans can, almost instantaneously, judge the quality of an image or video based on the prior knowledge they have learnt through viewing millions of images/videos during their daily life. Therefore, one straightforward way for video quality assessment is the subjective testing, which requires that humans provide their opinions of the image/video quality. However, it is very time-consuming and expensive, which makes it impractical for video applications. These drawbacks lead to the development of video quality metrics which can automatically evaluate the video perceptual quality.

The simplest and most popularly used video quality metric is the Mean Squared Error (MSE) and related measures such as Peak Signal-to-Noise Ratio (PSNR), mostly due to their simple formulations, clear physical meanings and easy mathematical optimizations. However, it is well known that the MSE and PSNR do not correlate very well with visual quality [1]. That is why a great deal of efforts has been made to develop objective image and video quality metrics, which incorporate HVS perceptual characteristics. In [2]-[4], JND model is employed to evaluate image/video quality by considering visual contrast sensitivity, luminance adaptation, and video temporal characteristics. Recently, VQA techniques attempted to characterize the features that HVS may associate with loss of quality, such as blurring, blocking, sharpness and so on. The most popular Image Quality Assessment (IQA) and VQA algorithms that embody this approach include Structural SIMilarity (SSIM) index [5]-[7], Visual Information Fidelity (VIF) [9], and Video Quality Metric (VQM) [8]. SSIM tries to capture the structure information loss to depict the distorted image quality. VIF employs the mutual information between the original and test image to evaluate the image quality. As SSIM and VIF perform on images, we can extend them to videos by applying them frame-by-frame, and the video quality index is obtained by averaging the frame level quality scores. Due to this strategy, SSIM and VIF fail to capture the temporal distortions between adjacent video frames. VQM, which has been adopted by the American National Standards Institute (ANSI) as a national standard for its excellent performance, analyzes 3D spatio-temporal blocks of the video sequence in order to extract the salient features, and the feature differences are employed to generate the video quality index. However, only the frame differences are involved in the VQM temporal component, which is too simple to depict the HVS temporal property. In order to integrate the explicit motion information, MOtion-based Video Integrity Evaluation index (MOVIE) [10] is proposed by tracking perceptually relevant distortions along motion trajectories, thus augmenting the measurement of spatial artifacts in videos. However, it emphasizes on the spatial and motion distortions, while ignores the HVS responses of the distortions. Moreover, it is very complicated for practical applications.

Recently, ABT-JND [11] for videos has shown its potential for modeling the HVS characteristics by not only considering the motion information of object and eye movement, but also accounting the temporal contrast sensitivity function (CSF). Furthermore, ABT-JND characterizes the HVS properties over different block-size transforms. Therefore, by incorporating the ABT-JND model into VQA algorithm, a good performance could be expected, for its accurate model of the HVS property.

This work was partially supported by a grant from the Chinese University of Hong Kong under the Focused Investment Scheme (Project 1903003).

978-1-4244-7993-1/10/$26.00 ©2010 IEEE

In this paper, a novel VQA based on ABT-JND is proposed to assess video quality by exploiting HVS characteristics of the spatio-temporal distortions over different block-size transforms. Firstly, ABT-JND model is applied on the original video sequence to generate its JND map and the corresponding macroblock type. Secondly, by referring to the obtained macroblock type, the differential frames between original and the distorted ones are transformed from spatial to frequency domain, specifically (Discrete Cosine Transform) DCT domain, by using different block-size DCTs. Then the frame-level quality score is obtained by accumulating the HVS responses of distortions based on the ABT-JND map. Finally, the video quality index is obtained by averaging the frame-level scores.

The rest of the paper is organized as follows. In Section 2, the ABT-JND model for videos is described. Our proposed VQA is introduced in Section 3. And experimental results are demonstrated in Section 4. Finally, Section 5 concludes the paper.

2. ABT-JND MODEL FOR VIDEO

Video JND profile in the DCT domain could be determined by considering both spatial JND profile T_{spatio} and temporal modulation factor T_{tempo}, which is defined as:

$$T(f,typ,m,n,i,j)=T_{spatio}(typ,m,n,i,j)\times T_{tempo}(f,typ,m,n,i,j)\,, \quad (1)$$

where f denotes the frame index of the video sequence, (m,n) is the position of macroblock in the current frame, (i,j) indicates the DCT subband, typ denotes the final transform size to generate the resulting JND map T for each corresponding frame. As in [2] [12], the spatial JND profile T_{spatio} is determined by a basic visibility threshold T_{basic} generated from the spatial CSF, the luminance adaptation α_{lum} and contrast masking α_{cm}:

$$T_{spatio}(typ,m,n,i,j)=T_{basic}(typ,i,j)\times\alpha_{lum}(typ,m,n)\times\alpha_{cm}(typ,m,n,i,j)\,. \quad (2)$$

Furthermore, the temporal modulation factor T_{tempo} could be derived by considering the temporal CSF [11] [12], which is the function of temporal frequencies. From [13], the temporal frequency of a video signal depends on not only the object motion information, but also its spatial frequency.

In the developed ABT-JND model [11], the motion information of the object is approximated by the Block-based Motion Estimation (BME). Therefore, during the derivation of ABT-JND model for videos, the temporal information has been considered, as well as the spatial information. And the block-level BME could efficiently depict the temporal motion information, which appears much more accurately than the frame differences employed in VQM. Furthermore, as shown in [11], HVS performs differently over different block-size DCTs. Specifically, larger block-size DCTs could provide better energy compaction and better preservation of details, while smaller ones could prevent more ringing artifacts during compression [14]. By introducing ABT-JND into VQA, HVS responses of the distortions could be accurately modeled, which will result in a better performance.

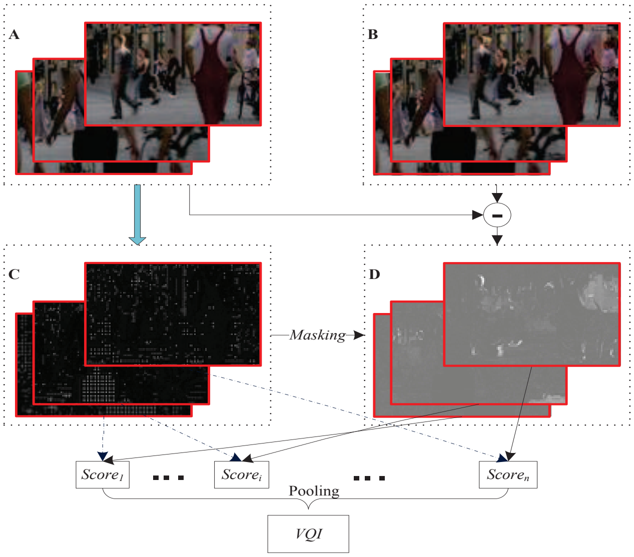


Figure 1. Proposed VQA framework based on ABT-JND.

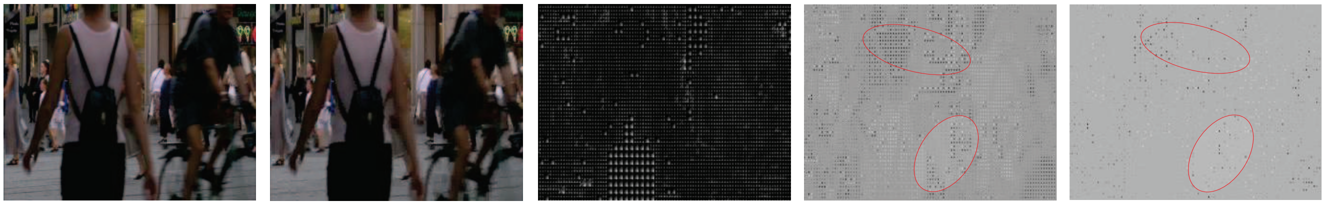
3. PROPOSED ABT-JND BASED VQA

The framework of ABT-JND based VQA is illustrated in Figure 1, which comprises several components: the original video sequence in block **A**, the distorted video sequence in block **B**, the differential video sequence in block **D**, and the JND maps generated from the original video sequence in block **C** for depicting HVS characteristics.

The video frames in the block **A** are the original ones without any distortions. Based on the original video sequence, HVS property of perceiving the sequence is modeled by ABT-JND, according to (1), which generates the corresponding JND map sequence in the block **C,** with consideration of both spatial and temporal HVS property. The distorted frames in the block **B** are generated by typical video distortions, such as: compression, transmission and so on. Then by referring to the original video sequence, the differential frames in block **D** are obtained by simply subtracting the distorted ones from the original ones, which is defined as:

$$D_{ist}(f)=O_{ri}(f)-I_{dist}(f)\,, \quad (3)$$

where $D_{ist}(f)$ is the *f-th* differential image between the original frame $O_{ri}(f)$ and the test distorted frame $I_{dist}(f)$, which are illustrated in Figure 2 (a) and (b). Moreover, according to the ABT-JND model introduced in Section 2, the JND map $T(typ,f,m,n,i,j)$ for the *f-th* original frame is also generated, by considering the local spatial content information, temporal motion characteristics, and different block-size DCTs [11] ,which is shown in Figure 2 (c). From the obtained JND map, we can see that all the sub-blocks (8×8) with the same content and similar movement are regarded as a unit, which employs 16×16 DCT-based JND model. However, the sub-blocks with diverse contents and irregular movement are processed individually, which means that 8×8 DCT-based JND model is utilized for each sub-block. As ABT-JND model generated in frequency domain tends to depict HVS



(a)Original frame O_{ri} (b)Test distorted frame I_{dist} (c)JND map T generated from O_{ri} (d)Frequency domain distortion DCT_{dist} (e)HVS responses P_{dist}

Figure 2. Processing results of the proposed VQA (Taking the 31-*th* frame of PA sequence in LIVE database as an example. (d) and (e) have been enhanced for better visualization with the identical processing).

responses of different frequency components, the differential frames are transformed to DCT domain, by referring to the block-size information obtained during the ABT-JND model derivation, which is depicted in Figure 2 (d). We can see that some regions, highlighted by red ellipses, represent larger absolute differences. However, HVS may not be sensitive to these differences compared with the others. Consequently, ABT-JND is employed to model HVS responses for masking the corresponding distortions, which is defined according to:

$$P_{dist}(f,typ,m,n,i,j)=\tau_{typ}\cdot\frac{DCT_{dist}(f,typ,m,n,i,j)}{T(f,typ,m,n,i,j)},\qquad(4)$$

where DCT_{dist} denotes the coefficient generated from $D_{ist}(f)$ according to different block-size DCTs, which are consistent with the ones for obtaining ABT-JND model. And the adjustable parameter τ_{typ} is introduced according to the different energy compaction and detailed information preservation properties of different block-size DCTs. Therefore, HVS responses of the distortions are approximately generated based upon the simple weighting function in (4), which is illustrated in Figure 2 (e). The image appears quite differently with Figure 2 (d), especially the part denoted by the red ellipses, which could more accurately depict the sensitive distortion that HVS perceived. Therefore, based on the HVS perceived distortion, a more convincing video quality metric can be derived.

As the video sequence is displayed and processed frame-by-frame, one frame quality score will be generated for each distorted frame, as shown in Figure 1. Since the HVS responses P_{dist} have been generated, a simple formulation like PSNR is employed for accumulating the HVS responses to generate the frame quality score, which is defined as:

$$\begin{aligned}&P(f)=\sum\nolimits_{(typ,m,n,i,j)}P_{dist}^{2}(f,typ,m,n,i,j)\\&Index_{p}(f)=10\log_{10}(MAX^{2}/P(f))\end{aligned},\qquad(5)$$

where P is the accumulated perceptual distortion which relates with the HVS responses, and MAX denotes the maximum HVS response of the distortion, which is always set as a constant for simplicity. As the quality score for each frame $Index_p$ has been generated according to (5), the Video Quality Index (VQI) of the specific distorted sequence is obtained by pooling all the frame quality scores together. In our implementation, the averaging process is employed to generate the VQI for simplicity:

$$VQI=\sum\nolimits_{f=1}^{N}Index_{p}(f)\big/N\,,\qquad(6)$$

where N is the total frame number of the video sequence. According to the definition of VQI, the higher the VQI, the better visual quality of the distorted sequence. And the VQI for the original sequence is infinite according to its definition.

4. EXPERIMENTAL RESULTS

In this section, we compare the performance of the proposed VQA method with the other VQA methods, i.e., PSNR, SSIM [6], Multi-scale SSIM (MSSIM) [7], VIF [9], VQM [8], and MOVIE [10]. We test all of the VQA methods on the LIVE video quality database [15] [16], which contains 150 distorted videos (obtained from 10 uncompressed reference videos of natural scenes). The distorted videos are created using four different commonly encountered distortion types, including MPEG-2 compression, H.264 compression, wireless distortions, and IP distortions. And each video sequence is evaluated by 38 human subjects for providing the Difference Mean Opinion Scores (DMOS).

In order to remove the nonlinearity, which is introduced by the subjective rating process, and to further facilitate empirical comparisons of different VQA methods, we follow the performance evaluation procedure employed in the Video Quality Experts Group (VQEG) HDTV test [17] and that in [18]. Let x_j represents the quality index that a VQA algorithm predicts for the *j-th* video sequence. A five parameter monotonic logistic function was employed for nonlinearly regression:

$$\begin{aligned}&VQI_{j}=\beta_{1}\log istic(\beta_{2},(x_{j}-\beta_{3}))+\beta_{4}x_{j}+\beta_{5}\\&\log istic(\tau,x)=1/2-1/(1+\exp(\tau x))\end{aligned}.\qquad(7)$$

After the nonlinearly mapping, the Pearson Linear Correlation Coefficients (LCC) between the objective and subjective scores, which measures the prediction accuracy, and Spearman Rank-Order Correlation Coefficients (SROCC), which measures the prediction monotonicity of the objective model prediction with respect to subjective scores, are employed to evaluate different IQA performances. Also the Root Mean Square prediction Error (RMSE) of the fitting procedure is also utilized to measure the VQAs' efficiency. According to the definitions, larger LCC and SROCC values mean that the objective and subjective scores correlate better, and smaller RMSE indicates a better performance.

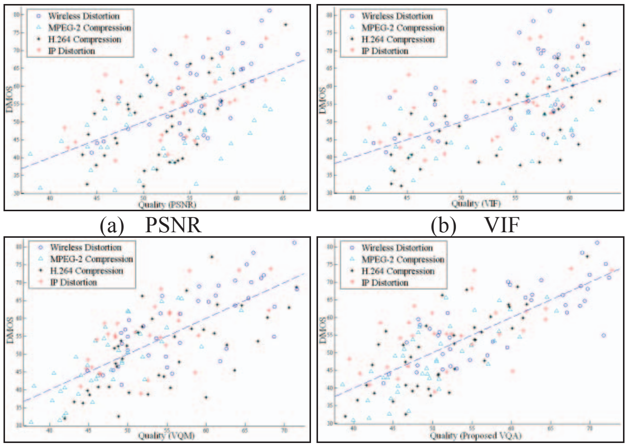
Table I. Performance comparisons between different VQAs.

VQA Methods	LCC	SROCC	RMSE
PSNR	0.5398	0.5234	9.241
SSIM	0.4999	0.5247	9.507
MSSIM	0.6754	0.7329	8.095
VIF	0.5735	0.5564	8.992
VQM	0.7160	0.7029	7.664
MOVIE[1]	0.8116	0.7890	-
Proposed	0.7627	0.7372	7.099

Detailed VQA comparison results are listed in Table I, which has shown that the proposed VQA outperforms the other VQA methods with larger LCC and SROCC values, and smaller RMSE value, while slightly poorer than MOVIE. Furthermore, the scatter-plots of different VQAs are shown in Figure 3. Intuitively, we can find that the spots of the proposed VQA scatter more closely around the fitted line than the other VQAs, which indicates a better performance of our method. Actually, although VIF and SSIM could perfectly depict the image visual quality, they fail to capture HVS responses of the temporal distortions, which lead to the poor performances. For VQM has considered temporal effect, it outperforms the typical IQAs, such as SSIM, PSNR, VIF. However, the temporal effect is just simply modeled by frame differences. Therefore, it could not efficiently depict the temporal distortions, which leads to a slight poor result. MOVIE is developed by considering the complex temporal and spatial distortion, which makes it perform the best. However, it is very complex and time-consuming. For each distorted sequence, the quality evaluation will take several hours, which is impractical for applications. As our proposed VQA has modeled the temporal HVS property as well as the spatial property, it provides a good performance. However, as ABT-JND is generated based on original sequence, it could be obtained offline. Therefore during video applications, simple weighting and summing computations are just employed to evaluate the test distorted video quality.

5. CONCLUSION

In this paper, we present a simple yet efficient VQA method, which considers not only the spatial and temporal HVS properties, but also the HVS responses over different block-size transforms. With evaluation on LIVE video subjective database, our proposed method outperforms other VQAs, while slightly poorer than MOVIE. Due to its simplicity, the proposed VQA could be easily applied on HVS-related video processing, such as video coding, video restoration and so on.



(a) PSNR (b) VIF

(c) VQM (d) Proposed VQA

Figure 3. Scatter-plots of different VQA methods.

6. REFERENCES

[1] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, MIT Press, 1993, pp. 207-220.

[2] W. Lin, L. Dong and P. Xue, "Visual Distortion Gauge Based on Discrimination of Noticeable Contrast Changes", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 900-909, July, 2005.

[3] A. B. Watson, J. Hu, and J. F. McGowan III, "Digital Video Quality Metric Based on Human Vision," *Journal Electronic Imaging*, vol. 10, no. 1, pp. 20-29, Jan. 2001.

[4] M. Masry, S. S. Hemami, and Y. Sermadevi, "A Scalable Wavelet-Based Video Distortion Metric and Applications," *IEEE Trans. Circuits Syst. Technol.*, vol. 16, no. 2, pp. 260-273, 2006.

[5] Z. Wang, L. Lu, and A. C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement," *Signal Process. Image Communication*, vol. 19, no. 2, pp.121-132, Feb. 2004.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vo. 13, no. 4, pp. 600-612, 2004.

[7] Z. Wang, E. Simoncelli, A. C. Bovik, and M. Matthews, "Multiscale Structural Similarity for Image Quality Assessment," *IEEE Asilomar Conf. Signals, System and Computers*, 2003.

[8] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Trans. Broadcast.,* vol. 50, no. 3, pp. 312-322, Sep. 2004.

[9] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," *IEEE Trans. Image Process.*, vol. 15. pp. 430-444, Feb. 2006.

[10] K. Seshadrinathan and A. C. Bovik, "Motion Tuned Spatio-temporal Quality Assessment of Natural Videos," *IEEE Trans. Image Process.*, Vol. 19, no. 2, pp. 335-350, Feb. 2010.

[11] L. Ma and King N. Ngan, "Adaptive Block-Size Transform Based Just-Noticeable Difference Profile for Videos", *ISCAS 2010*.

[12] Z. Wei, King N. Ngan, "Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Images/Video in DCT Domain," *IEEE Trans. Circuits Syst. Technol.*, vol. 19, no. 3, pp. 337-346, Mar. 2009.

[13] Y. Wang, J. Ostermann, and Y. Zhang, *Video Processing and Communications*, Prentice Hall, 2002.

[14] J. Dong, J, Lou, C. Zhang, and L. Yu, "A New Approach to Compatible Adaptive Block-Size Transforms," VCIP, 2005.

[15] K. Seshadrinathan, et al., "Study of Subjective and Objective Quality Assessment of Video", *IEEE Trans. Image Process*., Vol. 19, no. 6, pp. 1427-1441, Jun. 2010.

[16] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "A Subjective Study to Evaluate Video Quality Assessment Algorithms," *SPIE Human Vision and Electronic Imaging*, Jan. 2010.

[17] VQEG, Final report from the Video Quality Experts Group on the validation of Objective Models of Video Quality Assessment. [online] Available: http://ww.vqeg.org.

[18] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Trans. Image Process*. Vol. 15, pp. 3441-3452, 2006.

[1] LCC and SROCC values of MOVIE are obtained directly from [15], which does not provide the RMSE value.